9-2015

TOWARD A TWO-STAGE MODEL OF FREE CATEGORIZATION

Gregory J. Smith
California State University - San Bernardino, smitg306@coyote.csusb.edu

TOWARD A TWO-STAGE MODEL OF FREE CATEGORIZATION

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

General Experimental Psychology

---

by

Gregory James Smith

September 2015

TOWARD A TWO-STAGE MODEL OF FREE CATEGORIZATION

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Gregory James Smith

September 2015

Approved by:

Dr. John Clapper, Committee Chair, Psychology

Dr. Hideya Koshino, Committee Member

Dr. Robert Ricco, Committee Member

ABSTRACT

This research examines how comparison of objects underlies free categorization, an essential component of human cognition. Previous results using our binomial labeling task have shown that classification probabilities are affected in a graded manner as a function of similarity, i.e., the number of features shared by two objects. In a similarity rating task, people also rated objects sharing more features as more similar. However, the effect of matching features was approximately linear in the similarity task, but superadditive (exponential) in the labeling task. We hypothesize that this difference is due to the fact that people must select specific objects to compare prior to deciding whether to put them in the same category in the labeling task, while they were given specific pairs to compare in the rating task. Thus, the number of features shared by two objects could affect both stages (selection and comparison) in the labeling task, which might explain their super-additive effect, whereas it affected only the latter comparison stage in the similarity rating task. In this experiment, participants saw visual displays consisting of 16 objects from three novel superordinate artificial categories, and were asked to generate binomial (letter-number) labels for each object to indicate their super-and-subordinate category membership. Only one object could be viewed at a time, and these objects could be viewed in any order. This made it possible to record what objects people examine when labeling a given object, which in turn permits separate assessment of stage 1 (selection) versus stage 2 (comparison/decision). Our primary objective in this experiment was to determine whether the increase in

category labeling probabilities as a function of level of match (similarity) can be explained by increased sampling alone (stage 1 model), an increased perception of similarity following sampling (stage 2 model), or some combination (mixed model). The results were consistent with earlier studies in showing that the number of matching discrete features shared by two objects affected the probability of same-category label assignment. However, there was no effect of the level of match on the probability of visiting the first matching object while labeling the second. This suggests that the labeling effect is not due to differences in the likelihood of comparing matching objects (stage 1) as a function of the level of match. Thus, the present data provides support for a stage 2 only model, in which the evaluation of similarity is the primary component underlying the level of match effect on free categorization.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. John Clapper, whose guidance inspires a passion for the diligent pursuit of research and scholarship. Without his unwavering support this manuscript would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

LITERATURE REVIEW

Introduction

The ability to partition objects into categories is fundamental to human cognition, because categorization allows people to make adaptive inferences and predictions about the nature of their environments. One of the most central questions in the study of human categories concerns how such categories are acquired from experience. Traditionally, the field of categorization has distinguished between two very general types of category learning, defined by the presence or absence of feedback or direct intervention. In *supervised* learning, categories are acquired over successive trials under direct instruction from an external source. In such tasks, the learner attempts to classify a series of training instances from different categories, and receives corrective feedback on each trial. A parent teaching a child different types of animals using flashcards is an example of this type of learning. Each time a new animal is shown, the child attempts to name it, and the parent tells them if they are right or wrong. By attending to this feedback, the child eventually learns to place each animal into its appropriate category. By contrast, in *unsupervised* learning the learner is presented with a series of unlabeled training instances and no classification feedback is provided. Hence, in this type of task a person must discover any groups or categories for themselves.

The great majority of category learning research has focused on supervised learning, which is undoubtedly an important and powerful form of learning in the real world, as the above example of parental training illustrates. However, this form of category induction is limited to a finite set of circumstances, often an instructional or laboratory setting, in which accurate feedback is reliably available. As such, much of what is learned through research on supervised learning may not generalize to less constrained natural situations in which external guidance and feedback are absent. In the real world, categories are often acquired spontaneously through what would seem to be an artifact of moment-to-moment perception in the absence of third-party intervention, such as the passive observation of different types of foliage on a hillside or noticing that a flock of doves are distinct from a flock of pigeons (Clapper, 2015). Perceptual distinctions between objects sometimes occur automatically or spontaneously when viewing a scene without instruction or supervision. Unfamiliar or novel objects may appear to *pop out* from a scene and are readily distinguished from more familiar objects when they are encountered. This kind of unsupervised learning is arguably even more prevalent than supervised learning in everyday life, and is sometimes characterized as a more basic or natural process of category acquisition.

The general issue of the *naturalness* of category acquisition is arguably much more central to the research associated with unsupervised learning than that of supervised learning. In unsupervised learning, some form of shared structure is necessary for people to discover and create encompassing categories, while supervised learning permits the learning of completely arbitrary categories of objects that share nothing in common except a label. In fact, supervised learning experiments often use relatively artificial category structures that are only learned after a series of trials, which allow for gradual learning curves to be compared across conditions in order to test specific hypotheses about models of learning. However, in unsupervised learning experiments, the *learnability* of the category structures used is of considerable importance, such that if the objects used are too artificial or unnatural people may never detect categories on their own. Thus, the issue of what makes a psychologically *good* or *natural* category is a fundamental element in research on unsupervised learning.

Similarity as the Essence of Categorization

What, then, is a psychologically *good* or *natural* category? What do such highly natural or learnable categories look like? The prevailing view for several decades has been that objects belonging to particular categories are grouped by their overall similarity or *family resemblance*. Family resemblance formalizes the notion that items within a category have more shared features in common than items belonging to different categories, so that within-group similarity is maximized and between-group similarity minimized in categories with high family resemblance (Mervis & Rosch, 1981). As such, individual categories are perceptually distinguished by a non-random correlational distribution of features, without any single necessary or defining feature invariantly predicting category membership.

The assumption that similarity underlies categorization stems from the relatively intuitive notion that groups should incorporate common objects sharing properties that are alike. It would seem unreasonable that a given category should include items that are not bound by some consistent and overlapping structure. Rosch, Simpson, and Miller (1976) demonstrate that objects typifying a particular category prototype are more easily learned and classified more quickly than atypical objects, such that a raven would be more quickly learned and identified as belonging to a category of birds than would an ostrich, illustrating the notion that the learnability of a given set of objects is a reliable indicator of category goodness (Pothos & Chater, 2002). Family resemblance structure is considered particularly representative of objects at the basic level, at which the ratio of within-category to between-category similarity is maximized relative to more superordinate or subordinate classification levels (Mervis & Rosch, 1981). Tversky and Hemenway (1984) showed that there is considerable agreement about the shared structure that constitutes a good category, particularly at the basic level, suggesting that similarity based categories are not spurious or arbitrarily defined.

Categories are presumably used by humans to organize the objects and events encountered in everyday life, and should thus reflect a real measurable order found in the world. Normative models of category goodness, such as Anderson’s (1990, 1991) Rational Model of Categorization or Gluck and Corter's (1985) Category Utility model, characterize categories as being useful for making inferences and predictions about newly encountered objects conditional upon these objects being assigned to known categories. For categories to be useful in this way, the objects within those categories must share a large proportion of features in common; in general, the greater the degree of similarity or feature overlap within a category (relative to neighboring categories), the greater the predictive power obtained by knowing that category (i.e., the greater the category utility, see Gluck & Corter, 1985). A category of *shiny things*, while imaginable, would be too broad to provide meaningful and informative relations pertaining to objects belonging to that category, and have relatively low predictive utility. Categories based on single features or dimensions provide little or no predictive power in and of themselves. Rather, it is the fact that categories capture correlational structure in the world that makes them useful for making predictions about the world. Categories that are too narrow and do not sufficiently capture the essence of a particular set of objects are not a particularly valuable method of organization, and provide a very weak framework for making inferences about the nature of objects, necessary for acquiring new instances of a particular category.

Support for Family Resemblance or Holistic Similarity

If psychologically *natural* categories in the real world are organized using family resemblance structure, it is expected that the same kinds of category structure would be easily recognized in laboratory tasks, including free categorization tasks in which no feedback is provided. In the next section, we briefly review the pertinent free sorting, category construction, and match-to-samples literature that tests this prediction. Collectively, this literature provides surprisingly little evidence to support the recognition of family resemblance categories, at least for relatively complex, discretely-varying objects such as those typically found in nature.

Some experiments have provided data consistent with similarity-based categorization. For example, Pothos and Chater (2002) reported several experiments testing their *simplicity model*, which formalizes the notion that categories are created to maximize internal similarity and minimize external similarity as predicted by family resemblance theory. In a series of studies using clusters of points arranged in geometric patterns, Pothos and Chater showed that stimulus properties, task type, and instructions all influenced performance on pattern sorting. Density within a cluster and distance in relation to other clusters, as well as the degree of cluster integration, was manipulated over several studies using both free categorization, in which the goal of sorting was relatively abstract and undefined, as well as more contextualized tasks in which family resemblance was more explicitly defined as the target of categorization. Taken together, their results provided strong evidence for simplicity (similarity) based grouping, thus lending support to family resemblance as underlying categorization in a variety of conditions. Subsequent free-sorting experiments by Pothos et al. (2011) have replicated this preference for similarity (simplicity) based sorting using a variety of stimulus materials and category structures.

Triad tasks, in which three items are presented and two must be selected as belonging together, allow for a direct comparison between people's tendency to categorize on the basis of overall similarity as opposed to individual dimensions of the stimuli. Ward (1983) examined how response latency influenced the pairs that were selected in a triad task using triplets of dot stimuli, with individual stimuli varying along seven possible values of length and density. Within each triad, one pair matched exactly on a single dimension, while the other pair did not match on either dimension but had greater similarity overall. Sorts based on holistic similarity were more often observed in participants who spent less time observing the display before making a decision. Conversely, longer decision times tended to correlate with sorting based on a single shared dimension of the stimuli. These differences were also observed when comparing participants deemed impulsive to those classified as reflective based on the Matching Familiar Figures (MFF) test, a standard measure of impulsivity. More impulsive participants tended to rely more on overall similarity while more reflective participants relied more on analytically based (one dimensional) sorting. When external time constraints were imposed by the experimenter, more similarity-based sorting was observed and one-dimensional sorting was reduced. Taken together, these results suggest that sorting on the basis of individual dimensions requires a more analytic, time-consuming approach, while similarity-based sorting is less demanding and requires less time to execute.

In a related set of experiments, decks of cards with objects varying along two separable dimensions, color and size, were sorted based on either first impressions or meticulous decision making (Smith & Kemler Nelson, 1984). Participants were told to group two of the three objects in 32 triads within a designated interval of time. Those in the first-impression condition were told to classify objects without thinking about it, while those in the meticulous condition were told to take all the time they needed to decide which objects went together the best. First-impression sorts were more likely to be based on overall similarity using both dimensions while the meticulous condition produced more unidimensional sorting, thought due to a more analytical comparison process that enables the selection of individual dimensions. Increasing cognitive load by including a mental subtraction task in the classification procedure reduced the proportion of unidimensional sorting and increased that of similarity-based sorting, supporting previous findings suggesting that sorting by overall similarity is more efficient and easier to accomplish than deconstructing objects and sorting based on individual dimensions.

The research summarized above provides evidence supporting family resemblance theory, with objects consistently assigned to groups based on overall similarity. Pothos and Chater (2002) demonstrated a preference for sorting based on within-category similarity and between-category dissimilarity, consistent with family resemblance, and results from the triad task suggest that under some circumstances family resemblance sorting is easier and more natural than sorting based on a single dimension. However, one limitation of this research concerns the specific types of stimulus sets and category structures employed. In particular, all of the experiments described above, in both the free sorting and triad tasks, use relatively simple stimuli composed of no more than two dimensions, and these dimensions varied in a continuous or graded manner rather than qualitatively or discretely. While people may be able to integrate feature similarity across a small number of dimensions of continuous or metric variation, they may not be as proficient calculating holistic similarity for objects with more dimensions and when those dimensions vary discretely rather than continuously. This is an important limitation, as objects in nature are typically complex and often vary both discretely and continuously. For family resemblance theory to be relevant to real-world categorization, it must sufficiently characterize these more complex cases as well as the kinds of simple stimuli used in the experiments described above.

Evidence Against Family Resemblance or Holistic Similarity

In fact, a substantial body of research indicates that people do not always classify on the basis of overall similarity. In an early study, Handel and Imai (1972) showed that the analyzability of objects influences category outcomes and sorting strategies. In their experiment, objects varying along two dimensions were categorized in a free sorting paradigm. The dimensions of the objects varied in their degree of structural distinguishability or separability across different sets, with more distinguishable feature combinations considered analyzable (separable) and less distinguishable feature combinations considered unanalyzable (integral). Unanalyzable stimuli (e.g., color patches varying in hue and saturation) were generally categorized in terms of overall similarity, presumably due to the difficulty associated with visually separating such dimensions. Alternatively, analyzable stimuli constructed of separable dimensions (e.g., geometric forms that varied in shape and color) were classified using a single-dimensional approach more often than in terms of overall similarity. Handel and Imai suggested that separability facilitates the isolation and comparison of individual dimensions, while integral dimensions tend to be perceived as a unitary whole, limiting the tendency to categorize in terms of a single dimension.

Imai and Garner (1965) also provided evidence for the importance of individual dimensions in sorting certain types of highly separable stimuli. Utilizing a free classification paradigm, Imai and Garner investigated how sorting preferences are affected by the discriminability of different dimensions, that is, the degree to which the different values of a given dimension are perceptually distinguishable. Stimuli composed of three binary dimensions (dots varying in position, distance between dots, and orientation), with discriminability varied from low to high along four levels for each dimension, were presented individually on cards. Participants were instructed to sort decks of cards into two piles of equal size (a constrained sorting task), with different decks containing different levels of discriminability across each of the dimensions. A strong preference for attributes with greater discriminability, or greater differences between individual values of an attribute, was observed. Under time constraints, sorting was faster when based on more easily distinguishable attributes compared to less distinguishable attributes. Results indicate that people often prefer to sort on the basis of a single attribute or dimension, and that the preference for a given attribute depends on its' discriminability relative to that of potentially competing attributes.

One of the most widely-known and influential demonstrations of people's preference for one-dimensional as opposed to family resemblance sorting was a series of experiments carried out by Medin, Wattenmaker, and Hampson (1987). They operationalized family resemblance structure in a stimulus set composed of four discrete binary dimensions, with each category defined by similarity to a prototypical exemplar and with no single feature serving as a perfectly diagnostic indicator for category membership (see Table 1).

Table 1

Four Dimensional Family Resemblance (FR) Category Structure from Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. Cognitive Psychology, 19, 242-279.

Type	Category 0	Category 1
Prototype	0 0 0 0	1 1 1 1
Object 1	0 0 0 1	1 1 1 0
Object 2	0 0 1 0	1 1 0 1
Object 3	0 1 0 0	1 0 1 1
Object 4	1 0 0 0	0 1 1 1

Note. All objects in each category share three out of four dimensions with the prototypical category, while no single feature among the set of objects is diagnostic of category membership.

Utilizing a category construction task, in which objects presented simultaneously are sorted into a predetermined number of categories (two in this case), they attempted to determine whether people would sort the objects on the basis of overall similarity or family resemblance. Instead, a strong insensitivity to family resemblance was observed. Rather than categories being constructed based on overall similarity, as shown in Table 1, participants showed a strong preference for using a single feature or dimension as the basis for category membership, sorting the objects based on that dimension while ignoring all other features of the objects (and hence overall similarity). This result suggests that unidimensional (or 1D) sorting may be more intuitive and easier to accomplish than calculating overlap across multiple features to derive similarity as predicted by family resemblance theory, at least when the objects concerned are relatively complex and vary discretely, rather than continuously, on each dimension.

In general, free sorting tasks involving more discrete object variation and more complexity seem to be associated with an increased probability of categorization along a single dimension. In such cases, people demonstrate little or no responsiveness to family resemblance categories. Ahn and Medin (1992) carried out a number of studies in which people were again found to base their categories on a single defining feature while paying little attention to other features of the objects. Verbal protocols suggested that a single preferred salient dimension was deliberately selected, with minimal reliance and attention allocated to additional features when creating categories. Such results lend further support to the notion that people are not sensitive to overall family resemblance and tend to focus on a single distinctive or highly salient dimension that can be easily used to partition the objects into categories. Although protocols indicate that categorization was driven primarily by an individual defining features, descriptions did occasionally encompass other features. Ahn and Medin argued that while this may contribute to categories that appear to be family resemblance based under some conditions, this only occurs when the required number of categories cannot be created directly via 1D sorting (e.g., when the dimensions of the stimuli have more than two values, but people are required to sort those stimuli into exactly two categories).

Regehr and Brooks (1995) showed that the method by which objects are presented in a sorting task has a strong influence on the type of strategy used to organize them into categories. A strong preference for 1D categorization was observed across a wide variety of stimulus types when a standard sorting or category construction task was used. Regehr and Brooks argued that this unidimensional sorting preference was an *array effect* as a result of the simultaneous presentation of all the objects in a given set. They suggested that under such conditions, people find it difficult to track multiple dimensions across multiple objects and instead focus on tracking a single dimension at a time, making it is easier to sort those objects using a unidimensional approach as opposed to a similarity-based approach.

In a second series of experiments, Regehr and Brooks (1995) used a so-called *match to standards* task, in which two category prototypes were displayed and participants were then asked to sort objects sequentially into one of the two categories defined by these prototypes. In this modified sorting task, people were more likely to sort on the basis of the overall similarity of each object in relation to the two prototypes, rather than their match or mismatch on a single dimension. In this task, categorization is reduced to a sequence of individual comparisons between each presented object and its' category prototype, which may require less attentional capacity compared to tracking feature overlap across multiple items simultaneously. This pattern of sequential pairwise comparisons may make it easier to compare each pair of objects in terms of multiple dimensions, as required to sort them based on overall similarity or family resemblance.

Milton and Wills (2004) carried out a number of experiments using Regehr and Brooks' (1995) match-to-standards procedure, with comparable results when similar stimuli were used. As before, two category prototypes were displayed side-by-side while a deck of cards was sorted one card at a time into the two categories defined by these prototypes, or standards, with each card placed face down below the category prototype deemed most similar. Once again, objects were categorized more by family resemblance than unidimensionally, supporting Regehr and Brooks' earlier arguments that the tendency to use single dimensions in whole-array sorting is due to the difficulty and high attentional load required to track multiple dimensions across multiple objects under these conditions.

Given the rather simplistic nature of artificial stimuli used in their first experiment, Milton and Wills (2004) also tried using more complex or realistic stimuli in several follow-up studies. The attributes of these stimuli (pictures of fictitious butterflies) were more integrated and less visually separable than those used in their first experiment. Somewhat surprisingly, these objects with less separable features were actually grouped into unidimensional categories more often than the seemingly more separable objects used in their earlier experiment. These results indicate that how people divide objects into categories depends to some degree on type of stimulus materials used. Importantly, additional manipulations showed that time constraints and increased attentional load tended to reduce FR categorization and bolster unidimensional sorting in the match-to-samples task (Milton et al., 2008). This suggests that sorting on the basis of overall similarity detection requires more attentional resources and is more time consuming than simple 1D sorting, even under conditions that do in fact elicit similarity-based sorting.

It is also important to note that the match-to-standards procedure is not necessarily representative of free and spontaneous categorization. Generally only two categories are permitted in this task, and these are created by the experimenter (who sets up the two prototypes as category standards) rather than the participant. Thus, the central characteristic of truly free categorization, that participants create their own categories, is absent from this task. Moreover, because people are forced to compare each object individually to the category prototypes, categorization is reduced to a process of sequential pairwise comparisons between individual pairs of objects rather than scanning over a whole set. As such, the results garnered from this constrained sorting procedure may not generalize to less restricted situations that may be more consistent with how natural objects are encountered and categorized in the real world.

While there is some support for family resemblance category organization from studies using the match-to-standards procedure, it seems clear that overall similarity assessment is not consistently utilized as the basis of free categorization in the absence of a variety of necessary conditions. In particular, similarity-based sorting is not observed in whole-array tasks or even consistently in the match-to-samples task depending on a variety of circumstances, including details of stimulus construction, time constraints, and attentional load. The fact that FR sorting actually appears more difficult and to require greater processing capacity than 1D sorting, even in the match-to-samples task, seems particularly damaging to the notion of similarity as the natural basis for human categories. Moreover, one could argue that any evidence for similarity-based sorting in the match-to-standards task is irrelevant to free categorization *per se*, since people are never allowed to create their own categories in this task.

To summarize, despite the fact that objects in the real world are generally thought to be categorized in terms of overall similarity (e.g., Rosch & Mervis, 1975), and this is corroborated by laboratory studies using stimuli constructed from a small number of continuously-varying dimensions as described earlier, the category membership of objects composed of multiple, discretely-varying dimensions seems more likely to be determined by a single dimension. Indeed, one could argue that there has *never* been a convincing demonstration of similarity-based free categorization using complex, discretely-varying stimuli, as opposed to those that merely vary along one or two metric dimensions. The fact that FR sorting can be obtained in the match-to-standards task does not seriously contradict this conclusion, as this task is not truly an example of free categorization. Even in this constrained and somewhat contrived task, more time, effort, and attentional capacity is required to engage in similarity-based sorting than to engage in 1D sorting. Thus, it is difficult to argue that similarity is really the default or *natural* strategy in such tasks. Given that objects encountered in nature are typically complex and exhibit both discrete and metric variation, the failure of similarity-based categorization is of real concern for the assumption that family resemblance accurately describes categories in the real world.

Holistic Similarity - An Unreliable Determinate of Object Categorization

Given that previous experiments in free classification have failed to demonstrate FR similarity-based categorization, it is apparent that people are not as sensitive to overall similarity, or correlational structure, as predicted by the standard Roschian view. People seem to have considerable difficulty integrating across complex, separable, and discretely-varying dimensions to assess overall similarity in the absence of feedback. While this may suggest that similarity is in fact not the natural basis for human categories, another possibility is that the operationalization of similarity in terms a common set of binary dimensions, as in Medin et al. (1987), may not be representative of similarity-based categorization in the real world. Certainly the kind of similarity used as a basis for categories in those experiments is difficult for the casual viewer to discover. To illustrate, Figure 1 shows a set of objects organized by family resemblance structure as defined by Medin et al. (1987) and others (panel A), and in terms of a single dimension (panel B). Even if shown the FR structure in advance, as in Figure 1A, it is very difficult to recognize the FR categories in a mixed array (the reader can try covering up Figure 1A and sorting the objects in Figure 1C into the FR categories). However, recovering the 1D sort shown in Figure 1B is very easy. Given the sheer difficulty of FR based categorization, people may simply resort to a unidimensional organization strategy to satisfy the perceived demands of the sorting task.

Free Categorization by Alignment and Feature Match

As the type of similarity defined by Medin et al. (1987), Regehr and Brooks (1995), and others does not lead to recognizable categories as predicted by Roschian family resemblance, it seems natural to ask whether a different way of defining similarity could be more effective for creating learnable categories. Clapper (2015) demonstrated that people are strongly sensitive to the overall *alignability* of examples of a category in which concrete objects share the same overall structural configuration or body plan, with a one-to-one mapping between corresponding parts of different objects. Figure 2 shows a set of three obvious categories of objects that are easily distinguishable by their alignable structure, even though they share no identical surface features in common. Each category of objects differ on four discretely varying dimensions, with the number of potential values on a given dimension as high as six.

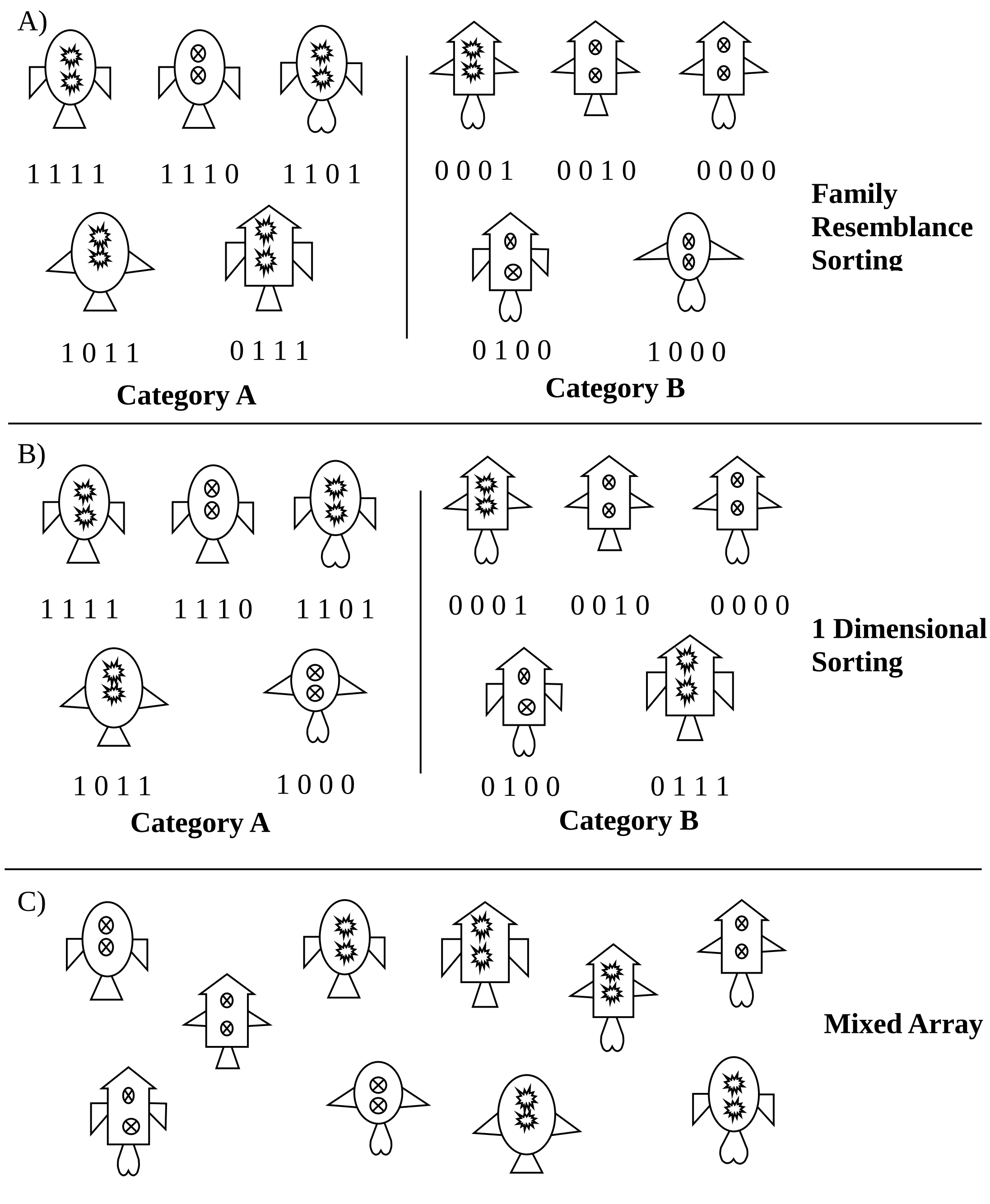


Figure 1. Two categories constructed of binary feature dimensions arranged by overall family resemblance versus a single dimension.

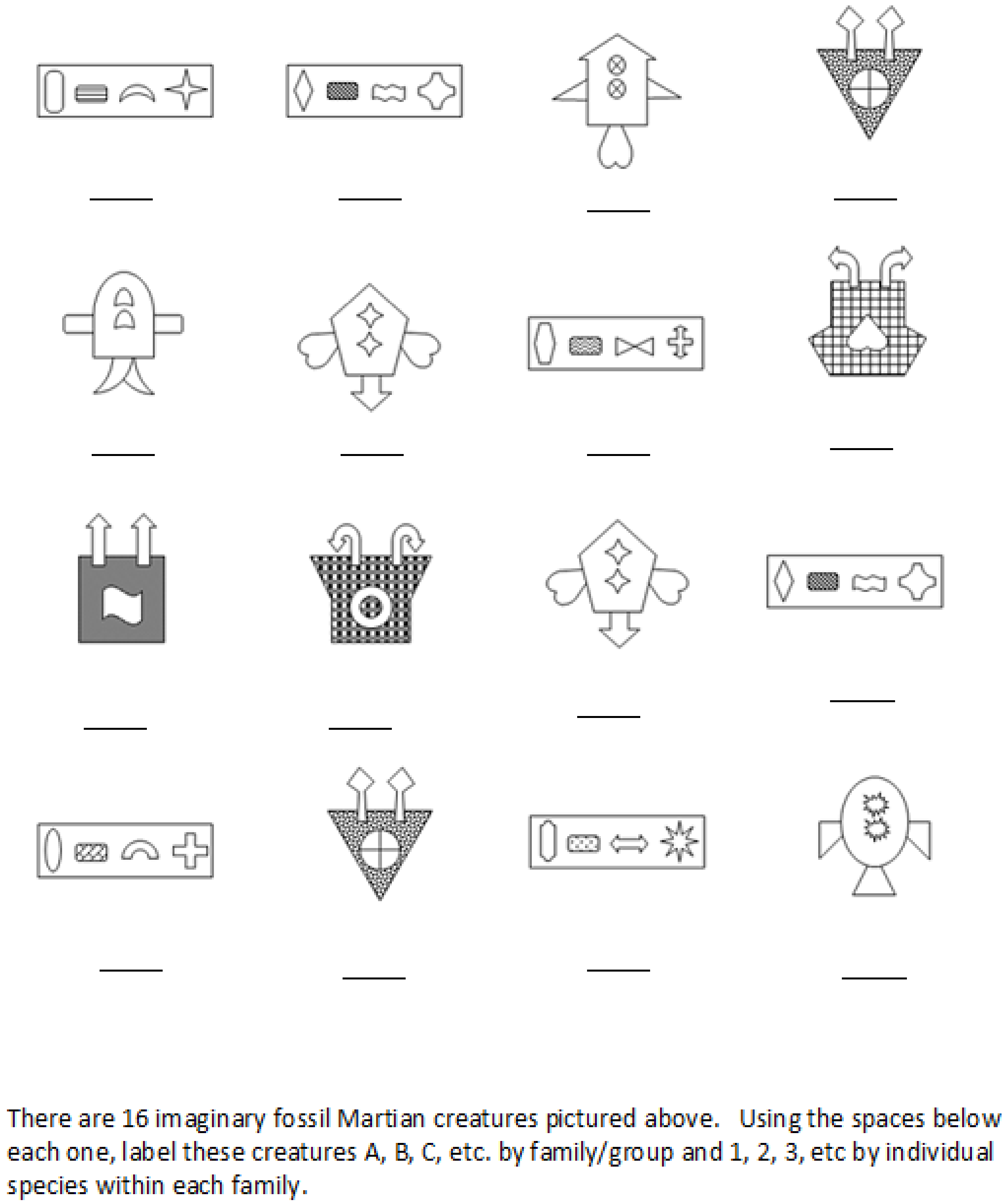


Figure 2. Illustration of three alignability-based categories arranged in a 4 x 4 display.

Clapper (2015) argued that this type of alignability-based category structure may be more representative of the type of stimulus domains encountered in nature than the type of FR structure used in previous research. For example, different birds share a similar overall structural configuration or body plan, with individually variable features such as beaks, wings, and talons maintaining a similar mapping across species. Note that in addition to superordinate categories based on overall alignability, a subordinate-level manipulation of category structure is also included in Figure 2. Specifically, an identical pair of matching instances is present within each alignable category, making it possible to investigate people's sensitivity to surface match as well as overall alignability.

This kind of stimulus set structure is inspired by models of similarity in which objects are assumed to be aligned before meaningful comparisons can be performed (Medin, Goldstone, & Gentner, 1993). The process of alignment is thought to occur rather spontaneously and subsequently facilitates a more effortful process of appraisal along individual alignable dimensions. Similarities and differences are more salient along corresponding alignable attributes than non-alignable attributes, making comparisons easier for alignable objects than for non-alignable objects. In general, alignable objects are perceived as similar overall and non-alignable as dissimilar (Clapper, 2015), with the degree of similarity then determined by evaluating the match between features along those structurally alignable dimensions.

Clapper (2015) aimed to determine whether participants would show sensitivity to this kind of alignability-based structure in a free categorization task.

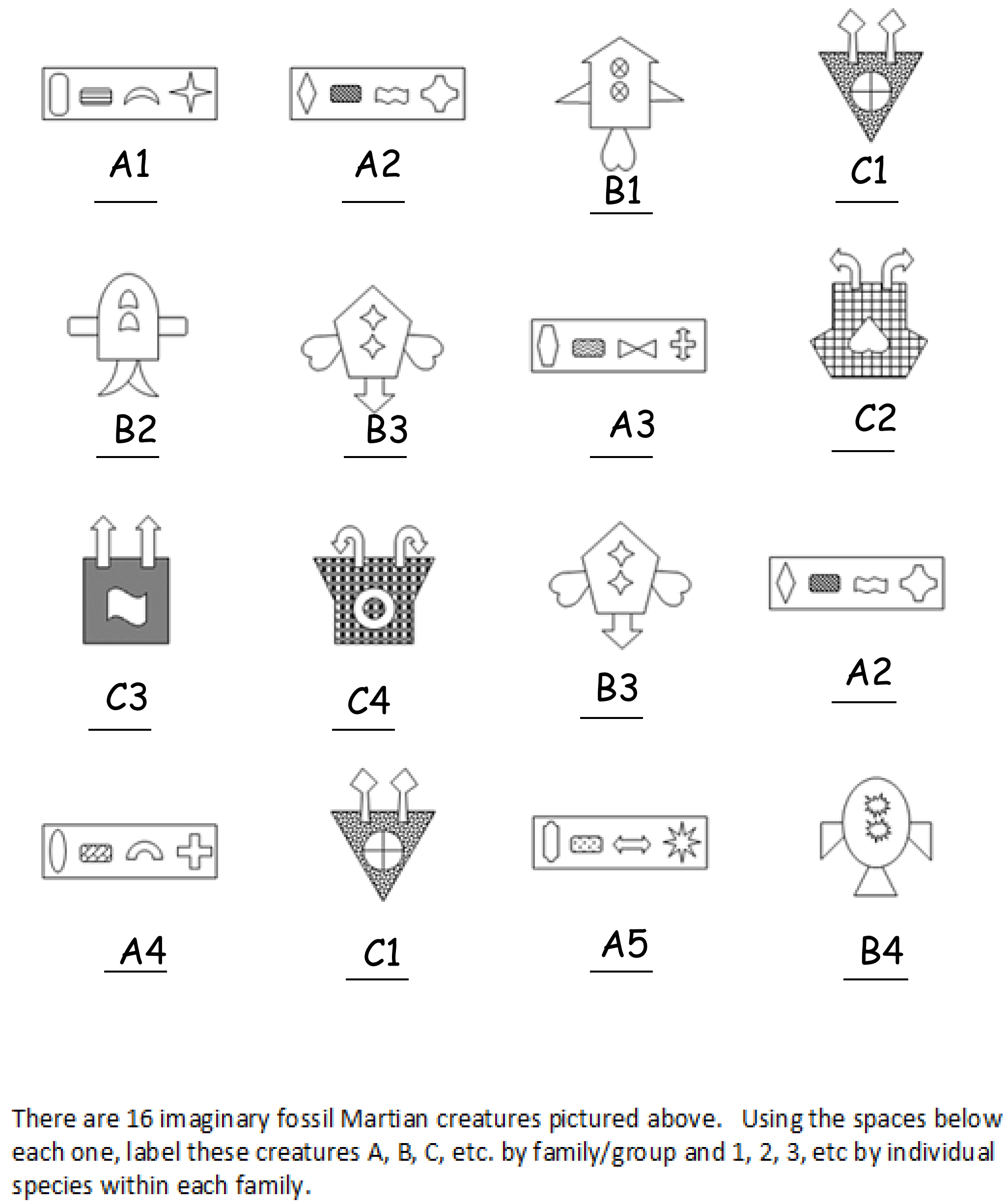


Figure 3. An example of a completed binomial labeling test page.

Rather than using a sorting task as in previous experiments, this research used a novel *binomial labeling* task in which two part letter-number labels (A1, B1, C1, etc.) were assigned to each object in an array, with letters representing overall families (superordinate categories) and numbers indicating species (subordinate categories) within those families (see Figure 3). As the goal was to determine the types of categories that are most psychologically natural and readily discovered, participants were not told how many families or species were present in a given display and were allowed to take as long as needed to complete the task to permit as much freedom of labeling as possible.

When presented with a stimulus set like that shown in Figure 2, people showed a strong sensitivity to both overall alignability as well as to feature-level identity within each alignable group. Thus, alignable objects were assigned the same family labels much more often than non-alignable objects, despite sharing no identical surface features in common (see Figure 3). In addition, matching instances within each alignable category were given the same species label significantly more often than other, non-identical objects from the same alignable category (Figure 3). Thus, people were sensitive to two kinds of similarity in this free categorization task, namely (1) the overall alignable or non-alignable structure between different objects and (2) individual discretely matching or mismatching features composing alignable objects.

These results are somewhat ambiguous regarding the type of surface match to which people were actually responding. The tendency to give matching objects the same species label might reflect sensitivity to individual discretely-matching features, or people might have simply recognized the overall identity of objects sharing four of four possible feature within an alignable category. This raises the question of whether people would put objects into the same species if they matched on some, but not all, surface features. And, if so, would the probability of same species labelling increase in a graded manner with the level of feature match?

Free Categorization as a Function of Matching Features

A pair of follow-up experiments by Clapper, Smith, and Miller (2015) investigated whether the number of matching features (discrete similarity) shared by two alignable objects would have a graded effect on the probability of species-level categorization in this task. As previously, three sets of broadly alignable objects constructed of four discretely varying attributes were arranged in a 4 x 4 grid.

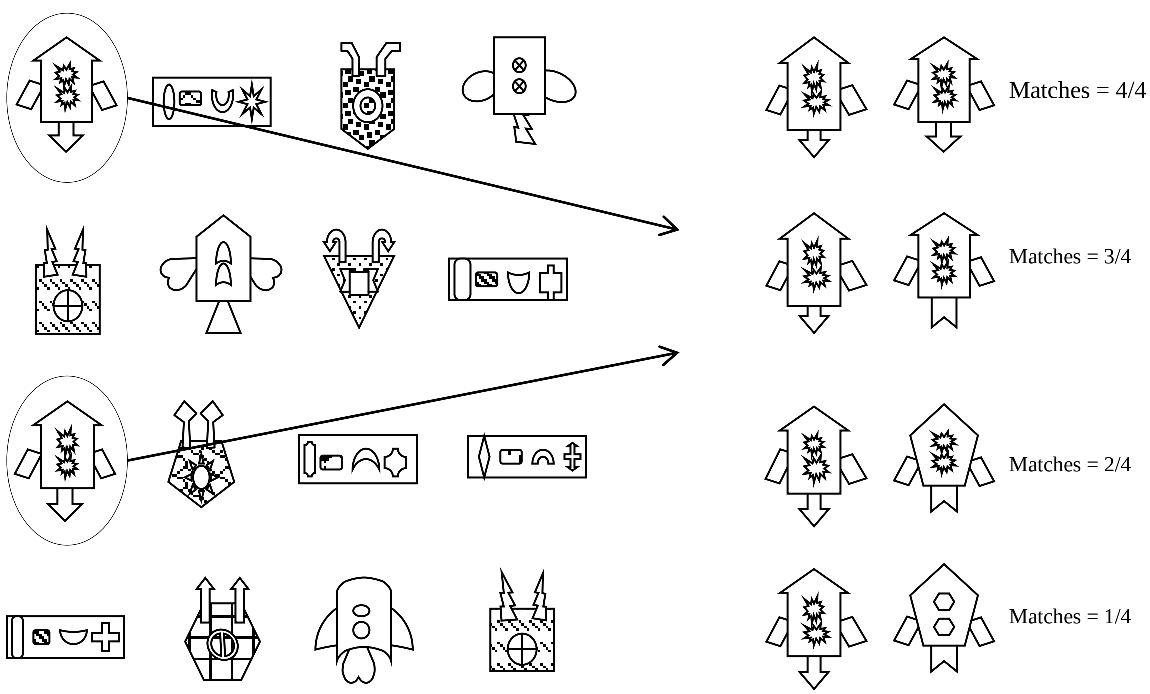


Figure 4. The level of match manipulation.

Matching instances were also included within each alignable family, with the number of identical features manipulated from one to four (out of four) to assess the influence of feature-level similarity on species labeling probabilities (see Figure 4). One of the experiments employed a between-groups design (manipulating levels of match across different groups) and the other a within-subjects design (manipulating levels of match within individual participants). Since their results were identical, these two experiments will be treated as one in the discussion below.

Note that this manipulation of within-group similarity is somewhat different from that used previous experiments, e.g., the FR structure of Medin et al. (1987) and others. The present stimuli are constructed of dimensions with more potential values than those of binary features. This means that the baseline probability that two objects will share a feature by chance is lower, making feature overlap have greater statistical relevance as compared to objects with fewer possible feature values. In the binary feature constructions used in previous categorization experiments, the probability of objects matching on a given feature is high even in the absence of correlational set structure, i.e., even if all dimensions vary independently and there are no real categories in the set. Hence, such matches should not be considered very informative when observed. When the dimensions of the objects have greater variance, sharing features in common should have a greater weight, particularly as the proportion of matching features increases, as multiple matches become exponentially more unlikely to occur by chance as their number increases.

Another issue with traditional FR structure is that features overlap between categories in these sets, making the categories difficult to distinguish. In the present experiments, the target species-level pairs share no features with other objects in their respective alignability-based categories, creating a situation in which within-category similarity can be manipulated in a graded manner while keeping the target objects clearly distinct from all non-target objects.

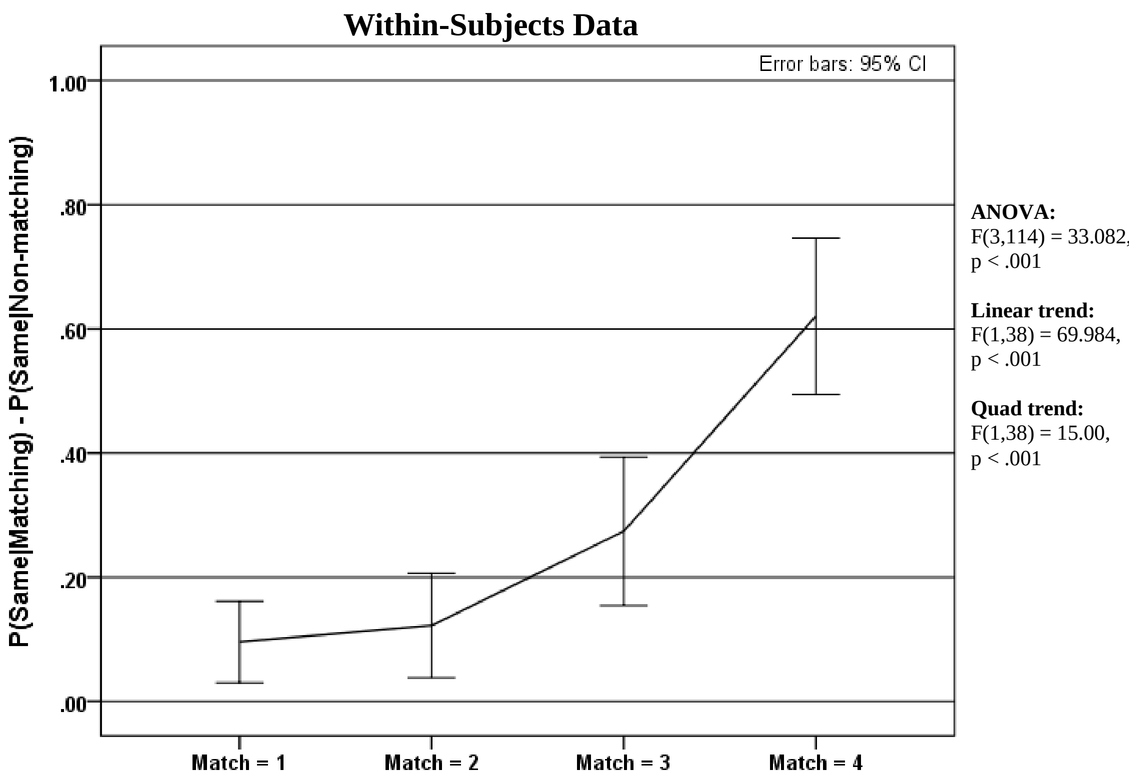


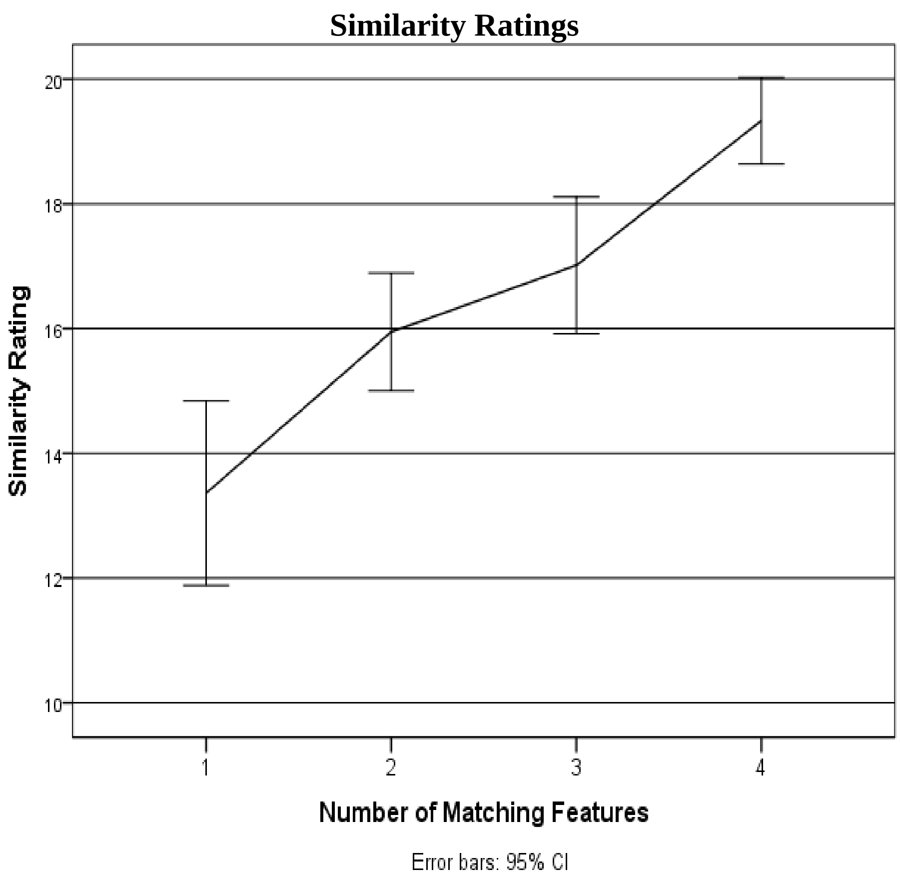
Figure 5. The probability of giving two objects the same species label, plotted as a function of the number of features they share.

This means that all that matters in the current task is whether people notice and consider relevant the matching features among the target objects, not whether they can distinguish relevant from irrelevant matches or track actual (statistical) correlational structure across the objects' features.

The usual pattern of alignability based superordinate categorization was observed, with objects from the same alignable categories being much more likely to receive the same family labels than objects from different alignable categories. At the species level, objects sharing a single identical feature were given the same label more often than other alignable objects without overlapping features, with each additional matching feature significantly increasing the probability of same-species classification (see Figure 5). The species labeling data showed a nonlinear trend as the degree of match increased from one to four (out of four), meaning that the impact of a matching feature increased with the overall degree of match. These results provide strong evidence that people are sensitive to manipulations of similarity via feature match within alignable groups.

In interpreting these results, it is natural to assume that manipulating shared features affects perceived similarity, which in turn affects the likelihood of species-level categorization. To directly assess the perceived similarity of objects used in the previous experiment, Clapper, Smith, and Miller (2015) presented participants with a 4 x 4 array with each object individually numbered from one to 16, and asked them to rate the similarity of selected pairs on a 20-point scale. As previously, three alignable families of objects were included in the display, and matching instances within each family again varied on their degree of match from one to four (out of four) features in common. Similarity judgments were collected between objects within and between different alignable families, and between pairs of objects within families sharing one to four identical features.

As expected, objects sharing the same alignable structure were rated as much more similar than objects from different alignable families.



Features Shared	Mean
1	13.36
2	15.95
3	17.02
4	19.33

Figure 6. Similarity ratings for directed comparisons plotted as a function of the number of shared features.

Similarity ratings of matching objects showed a roughly linear trend as a function of degree of match, with each additional shared feature increasing rated similarity by approximately the same amount (see Figure 6). The disparity between the superadditive function found in the categorization task and the linear function in the similarity rating task suggests that somewhat different factors must be at work in the two tasks.

A Two Stage Model of Categorization

Given that labeling versus rating tasks exhibit different sensitivity functions in response to levels of feature match, a unitary construct of similarity cannot account for both results. One way to explain this difference is to appeal to differences in the underlying processes by which similarity affects behavior in these two types of tasks. Considering what is required to assign the two matching objects the same species label in the unsupervised free categorization task, there are two steps, or stages, that must occur. First, an observer must compare and notice that a set of given objects share common features, what might be referred to as a *sampling* stage. From the 120 potential comparisons can be performed across a set of 16 objects, it seems obvious that some comparisons will be made while many others will not. Assuming a given pair is selected in stage one, an observer must then decide if there is some significant commonality that warrants the same category assignment, or whether the objects are not enough sufficiently alike and belong in separate categories. This might be referred to as an *evaluation* stage. The similarity of objects might affect their stage 1 sampling probability (i.e., the likelihood of comparing those objects in the first place), the stage 2 decision whether or not to assign them the same category label, or some combination of the two. Pairs comprised of more matching features may be more readily noticed and likely to be compared during stage 1, and subsequently more likely to be given the same category label in stage 2, than those sharing fewer identical dimensions. In the similarity rating task described above, people are told to compare particular pairs, thus eliminating stage 1 and leaving only the stage 2 evaluation relevant. This could explain why matching features have a progressively larger effect in the labeling task but not the direct rating task.

Two stage models that feature an automatic noticing, or retrieval, stage followed by a more deliberative judgment, or evaluation stage, are relatively common in the psychological literature. For example, Anderson's (1982) model of skilled performance stipulates that potential production rules or templates are retrieved automatically from procedural memory based on the available cues in a given task situation. After this, the best-matching rule is then selected by a slower and more controlled process that determines the most precisely matching option appropriate for the current situation. As another example, standard models of analogy (e.g., Gentner, 1983) assume that people notice potential analogies based on automatic memory retrieval due to surface match, while the goodness or quality of the potential analogy is then determined by a more effortful examination of the structural match between source and the target objects. The relation between the present model and other two-stage models will be explored at greater length in the General Discussion.

Testing the Model - Disentangling Sampling and Evaluation

The primary goal of this thesis experiment is to investigate the two-stage model by providing independent measures of sampling (stage 1) versus final categorization (stage 1 plus stage 2). The present experiment uses a novel variation of the binomial labeling task that enables direct and independent examination of the comparisons made across a set of objects as well as the labels subsequently generated. In this task, occluded objects from the same family-species structure as in the matching-feature experiments of Clapper (2015) are presented on a computer screen in a 4 x 4 array. By using the mouse to navigate the display, a single object can be viewed at a given time while all other objects in the display remain invisible. Labels are entered by typing them into response boxes located below each object in the display. Objects can be visited and/or labeled in any order, and may be revisited to make comparisons or modification to previously assigned labels. The main data of interest are the labels that participants enter for each object, as well as the pattern of comparisons they make with other objects as they do so.

In a preliminary experiment using this procedure, a consistent scanning pattern was often observed in which a new, to-be-labeled object would first be visited for an initial inspection, followed by a number of comparisons with, or visits to, other objects in the array before the initial object was finally assigned a label. For data analysis purposes, we referred to this interval of comparison (visits to other objects) preceding the final labeling of a given object as an *epoch*. Within a given epoch, there were significantly more comparisons with alignable objects than with non-alignable objects. In other words, at the superordinate level we found a greater proportion of within- than between-category comparisons within each epoch. Of particular interest was the pattern of comparisons people made as they were preparing to label the second member of a matching pair. Epochs for this instance revealed proportionally more comparisons to the first identical instance than to other alignable objects, or to objects from other, non-alignable categories. People were also significantly more likely to give the two matching objects the same species label, as in previous binomial labeling experiments discussed above. Of course, not everyone who compared the two matching objects gave them the same species label, so the probability of comparison (sampling) exceeded the probability of the two objects' receiving the same species label, which reflects both sampling and evaluation. As sampling patterns, or visits to other objects during a labeling epoch, were consistent with and predictive of the category labels generated, tracking such patterns would appear to be a promising way to provide an online measure of sampling or comparison (stage 1), which can then be related to the final pattern of category labels generated in a particular condition.

The current experiment is an extension of the earlier comparison tracking study with the inclusion of a within-group similarity manipulation similar to those used in some of the binomial labeling experiments described above. The number of features shared by matching instances within alignable families will be varied from one to four (out of four, see Figure 4) to assess the labeling pattern discussed above in light of the two stage model. In general, we expect the same results in the labeling component of this task as in previous experiments discussed above. Alignable objects should tend to receive the same family labels while matching objects within those alignable categories, or those with the greatest proportion of matching features, should be more likely to receive the same species labels. The probability of visiting alignable objects, non-alignable objects, the other matching object, and the target object itself will be compared across each level of feature match, with major interest focusing on visits to the other (first) matching object. The expectation is that the probability of visiting the first matching object, as well as the proportion of time spent visiting this object, will increase with levels of match.

Of particular interest in this experiment is whether the effects of similarity on the categories created is primarily due to changes in the sampling probability of the relevant pairs (stage 1), how these pairs are evaluated once sampled (stage 2), or some combination of both (see Figure 7). The interval of comparison, or visits to other objects, following the first visit and preceding the final labeling of a given object, referred to as an *epoch*, will be constrained to lengths that vary between three and 14, based on analysis of pilot data. An epoch length of less than three will necessarily have a same category visitation proportion of zero, .5, or one, while epochs longer than 15 consist of random sampling during which the initial object visited may no longer be the labeling target, or is forgotten and a new target object is assessed. If categorization is *conditional* on a stage 1 sampling-only model, the probability of visiting the first matching instance should increase with level of match, while the probability of assigning the same species label *conditional on visiting* should remain constant.

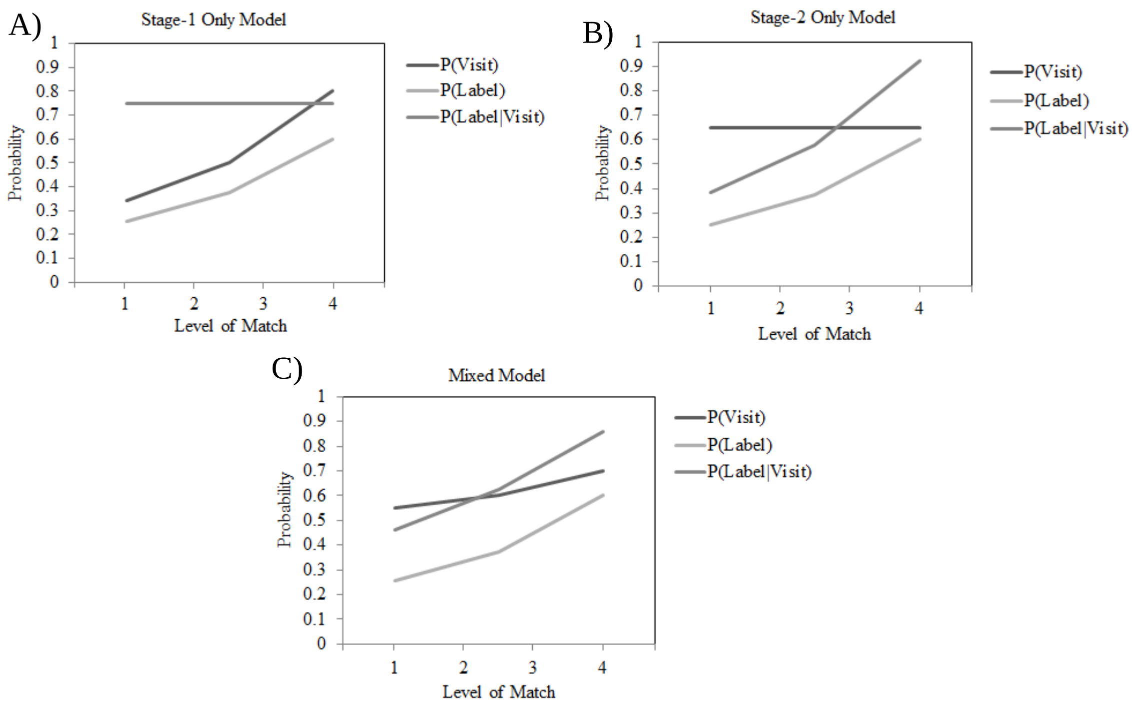


Figure 7. Possible patterns of results predicted by different models. In the stage 1 sampling-only model, the probability of visiting the first object P(visit) and the probability of assigning the same label to both objects P(label) both increase with levels of match, but P(label|visit) remains constant (panel A). In the stage 2 evaluation-only model, P(visit) remains constant while P(label) and P(label|visit) both increase with levels of match (panel B). In the mixed model, P(label), P(visit), and P(label|visit) all increase with levels of match (panel C).

In this case, the increase in sampling probability would explain all the increase in labeling probability as a function of match (Figure 7A).

If categorization is dependent solely on a stage 2 evaluation-only model, the probability of visiting (stage 1) should not vary as a function of level of match while the probability of assigning the same label should increase; here, all the change in labeling probability would be due to the stage 2 decision process (Figure 7B). However, if both the visitation and labeling probabilities are affected by the level of match, but the functions are not parallel -- in particular, if the conditional probability of labeling given sampling changes with level of match -- a mixed model would be supported, with the differences in labeling due in part to factors operating at both stages (Figure 7C). Thus, we not only want to determine whether levels of match affect co-labeling probabilities and the probability of visiting the first matching object when labeling the second, but also whether the effects of match are the same on both of these dependent variables. For the purposes of examining the effect of levels of match, the co-labeling probabilities for matching objects at each of the four levels of match are compared using a one-way ANOVA, with trend analyses and individual paired-samples t-test comparisons conducted for both family and species data, although the main focus for this effect is on the species-level data.

In addition to analyses based on the *probability* that people will or will not visit the first matching instance while labeling the second, the *proportion* of time spent viewing matching objects during a given epoch is also examined. In particular, given that a person visits the first matching instance at least once while labeling the second, does the proportion of time spent visiting that object vary as a function of levels of match, with objects with fewer matching features being examined for less time that those containing more matching features. And among those participants who do view the first matching instance, are those who spend more time examining it be more likely to assign the same species label to the second matching object.

CHAPER TWO

METHOD

Participants

Participants were 48 undergraduate students from the California State University, San Bernardino that received extra credit in a psychology course of their choice.

Procedure

Each participant was seated at one of twelve computers separated by cardboard dividers in a group lab. An informed consent page was placed in advance at each seat while instructions detailing the nature of the experiment were displayed on the monitor. For the labeling task, participants were asked to assume the role of an inter-planetary biologist tasked with classifying novel Martian fossils into broad families and identifying individual species within each family. They did this using the two part letter-number binomial labeling structure (A1, B1, C1, etc.) discussed above, with letters representing broad families (A, B, C, etc.) and numbers indicating individual species (1, 2, 3, etc.) within each family. A stem-plot showing natural examples of family and species relationships was included below the instructions for clarification (see Appendix A). Participants were prompted to use the spacebar to advance to a second page illustrating the task with 16 labeled images of unfamiliar plankton stimuli (from Haeckel's *Art Forms in Nature*, 1974) arranged in an example 4 x 4 display (see Appendix B). A third page showed the same 4 x 4 display with all images blurred (double 75% blur effect using GIMP Image Editor), except the object in the upper left corner, as would be seen during the start of the testing phase of the experiment (see Appendix C). Boxes designated for typed label responses were included below each image.

Once they were ready to begin the test phase, participants used the spacebar to advance to the first test display. Objects in the display were visible one at a time and revealed by using the mouse to navigate the display. Clicking on a desired occluded stimulus revealed the object in that position while subsequently blurring the previously visible image. The test phase consisted of a total of four 4 x 4 stimulus displays. An onscreen button was located at the bottom of the screen and used to advance to the next stimulus array once participants were finished labeling the 16 objects. Displays could not be revisited once moving on to the next page. After labeling the final stimulus array, participants were shown an instructional debriefing detailing the general purpose of the experiment, provided contact information should they have any questions, and thanked for their participation. The entire procedure took approximately 15 minutes to complete.

Materials and Design

The stimuli used during the test phase were constructed of four parts or components arranged in a consistent structural pattern, with a total of three different alignable structural arrangements defining different potential categories. Features were sourced from Microsoft Word shapes for the four dimensions of each object within an alignable category, with dimensions varying discretely along six possible values. Stimulus displays consisted of 16 objects shown in a 4 x 4 grid pattern, with a space provided underneath each object for participants' labels (see Appendix D). As noted, three structurally distinguishable alignable categories were used in each display. Three pairs of matching instances in each 4 x 4 array, one in each of the three alignable categories, shared one to four (out of four) identical features. Different matching pairs in the same display always had different levels of match. All other stimuli within each alignable category differed along each of the four dimensions, and none shared any values with other families or with either of the two matching stimuli.

Level of match was varied in a within-subjects design in the current experiment. Given that four levels of match are possible across the three matching pairs available in a given stimulus array (one for each alignability-based category), one of four possible combinations of match can be used in any given array (123, 124, 134, and 234). As such, each participant viewed four stimulus displays containing matching pair combinations exemplifying each of these four possible arrangements. Possible materials effects were controlled by creating four different stimulus sets, each composed of different objects from the three alignable categories constructed from different combinations of values. Each participant saw four slides, one from each of these sets. The order of array presentation was also counterbalanced to control for order effects.

To summarize, the primary independent variables in the current experiment concerned whether a given pair of objects in the same display (1) were alignable versus non-alignable; (2) for alignable objects, whether they were matching versus non-matching; and (3) for matching objects, whether they shared one, two, three, or four identical features. Dependent variables measured in the current experiments are as follows. First, the probability that a given pair of objects is assigned the same category label, coded at both family and species levels. Second, the probability and proportion of visits to other objects in the same alignable category, different alignable categories, the object itself, and to matching instances, if any, were recorded during the labeling epoch for each object in a given display.

CHAPTER THREE

RESULTS

Category Labeling Data

Each of the 16 two part family-species labels provided for each item in the four stimulus displays was coded into separate variables at both family and species levels. Each possible pair of objects from each display was coded in terms of whether the objects were given the same or different labels at each level. These same-different categorization matrices were the main dependent measure derived from the labeling data in this experiment, and were used to determine the probability of objects from the same versus different alignable categories being assigned the same labels, the probability of repeated matching instances within each alignable category being assigned the same label, and how the number of identical features shared by matching instances affected the probability that they received the same label. All comparisons were carried out at both family and species levels. Data from a total of 48 participants were collected for the current experiment, two of which were eliminated due to a failure to follow instructions, e.g., failure to complete the labeling task, or categorizing the objects only at a superordinate (family) level.

Sensitivity to Alignable Categories

Participant's sensitivity to the alignable categories was evaluated by comparing the probability of giving pairs of alignable objects the same label to that of giving non-alignable objects the same label. Since alignability was not fully crossed with the other factors in the design of this experiment, e.g., there were no non-alignable matching instances, this effect was evaluated with a t-test rather than ANOVA. We expected that alignable objects would be much more likely to receive the same label than non-alignable objects, especially at the family level.

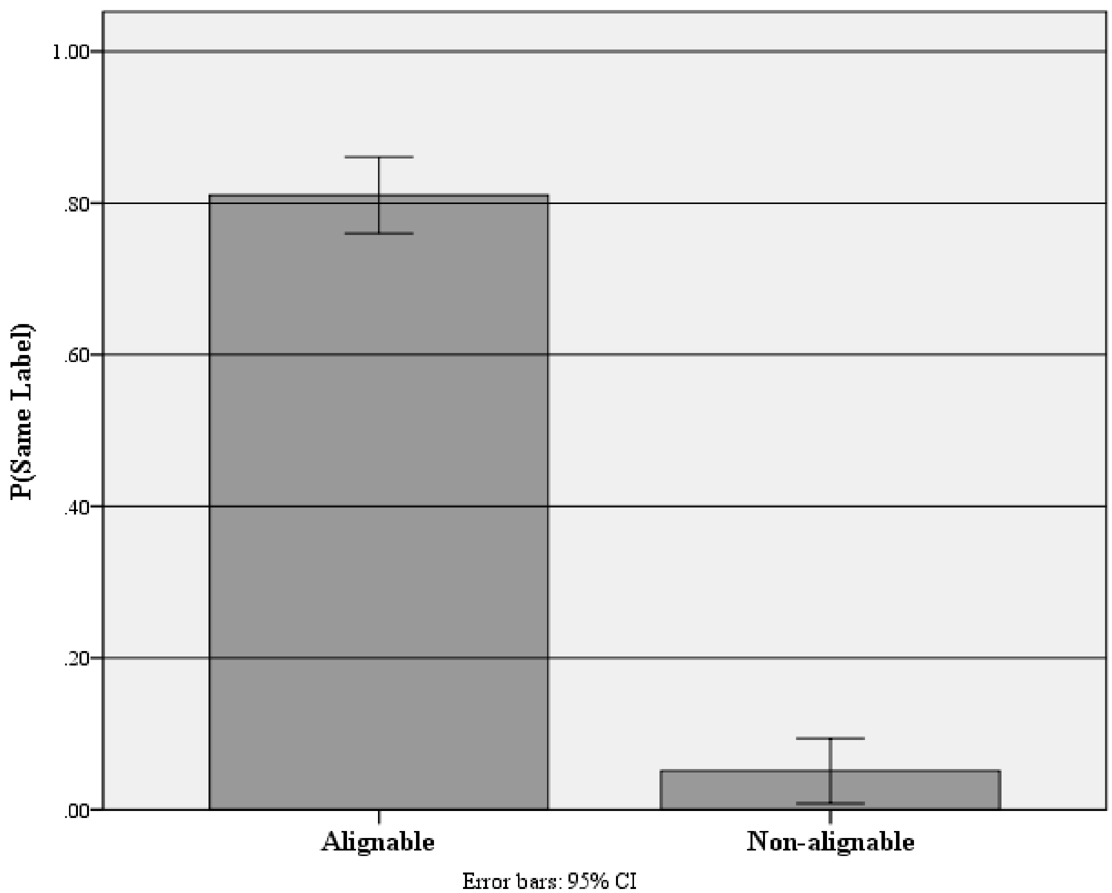


Figure 8. The probability of assigning the same family label to alignable versus non-alignable objects in a given array.

At the family level, alignable objects were given the same category label ($m = 0.81$) significantly more often than non-alignable objects ($m = 0.05$), see Figure 8, suggesting that people are indeed recognizing and utilizing structural alignability as a principle of categorization, $t(45) = 20.73$, $p < .001$. The same was true at the species level, with alignable objects again being categorized together ($m = 0.09$) more often than non-alignable objects ($m = 0.007$, $t(45) = 2.89$, $p < .01$. Note that the effect at the species level is less pronounced than that at the family level due to peoples' tendency to give each object its own unique species label). These results are consistent with previous results indicating a strong sensitivity to overall feature configuration (alignability), even when the objects share no identical surface features (i.e., specific attribute values).

<u>Sensitivity to Matching Features</u>

In addition to overall alignability, we also expected people to be sensitive to matching features in this experiment. In this second analysis, the independent variable is whether instances share matching features (i.e., matching vs. non-matching alignable objects). Again, this factor was not fully crossed with others in a factorial design (e.g., non-matching objects do not vary in level of match), so this effect was also evaluated with a t-test. Matching features were expected to mainly impact species-level categorization, so we will restrict our attention to the species data here. Overall, the species level data showed more grouping of matching ($m = 0.19$) than non-matching ($m = 0.09$) pairs within the same alignable family, $t(45) = 3.90$, $p < .001$, see Figure 9. However, our major goal in this experiment was to determine whether the size of this matching-feature effect varied with the number of matching features shared by two objects.

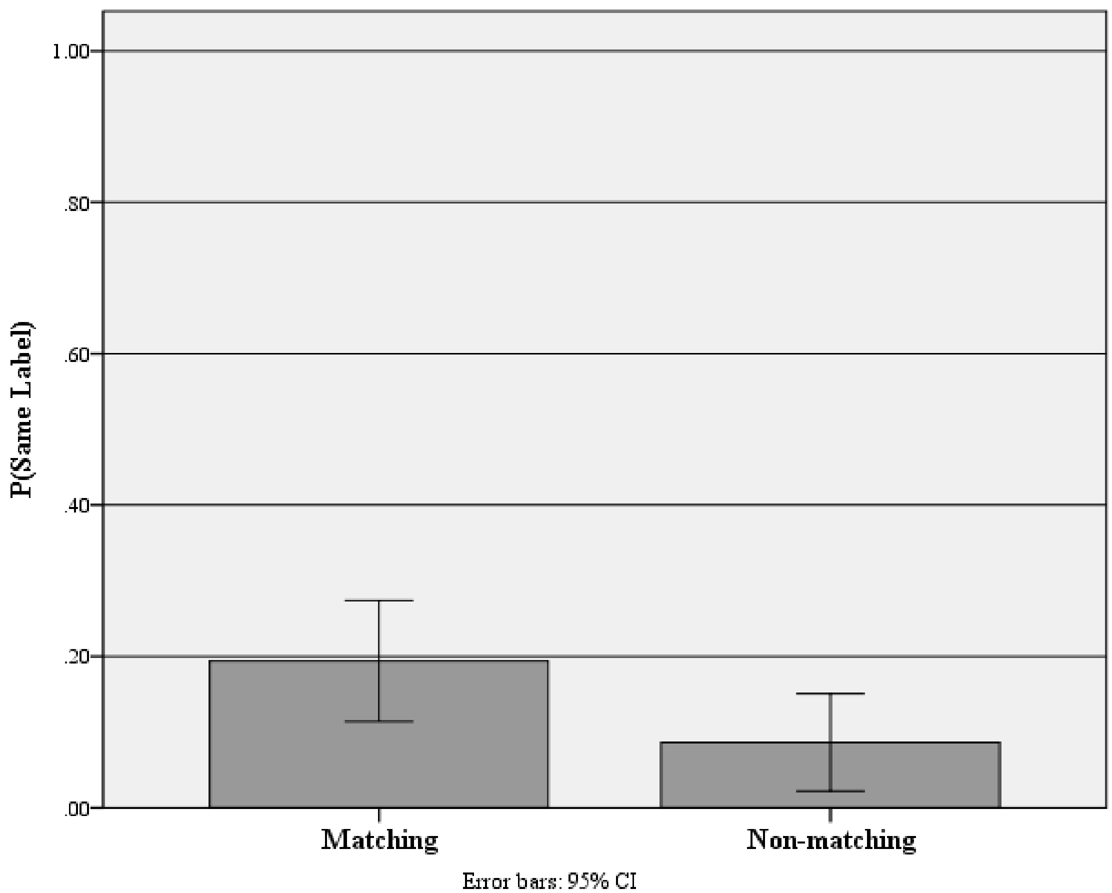


Figure 9. The probability of assigning the same species label to objects matching on at least one feature vs. alignable, but non-matching objects.

Here, we report analyses based on data from which the bottom 20 percent of participants have been excluded (n = 37), where performance was defined in terms of sensitivity to the alignable categories in the family-level labeling data. In other words, a *fit score* was computed for each participant, and this was used to rank their performance. This was done to improve the sensitivity of our analyses by removing inattentive or careless responders from the data.

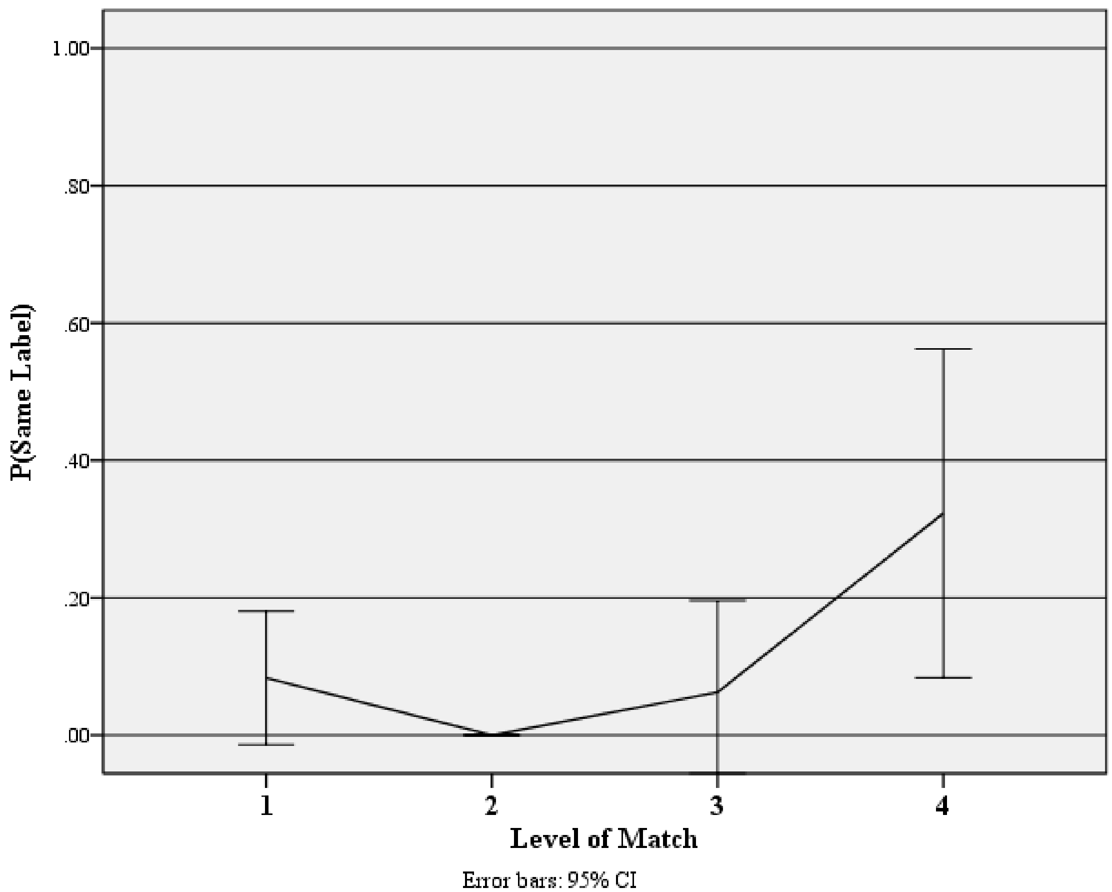


Figure 10. The probability of objects receiving the same species label as a function of level-of-match.

There was a significant effect of the level of match (one to four matching features) on the probability of same-label assignment for alignable objects, $F(3, 45) = 3.92$, $\eta_p^2 = .207$, $p < .05$[1]. Objects sharing four completely identical features (m = 0.33) were grouped together significantly more often than objects sharing three (m = 0.09, $p < .05$), two (m = 0.05, $p < .05$), or one identical feature (m = 0.11, $p < .05$), as can be seen in Figure 10 (all pairwise comparisons were made using paired-samples t-tests). There were no significant differences between labeling probabilities for objects sharing one to three matching features, $F(2, 38) = 0.72$, $\eta_p^2 = .037$, $p = .49$[2]. Curve estimation over all levels of match indicates the data fit both linear (marginally) and quadratic trends, $F(1, 15) = 4.07$, $\eta_p^2 = .213$, $p = .06$ and $F(1, 15) = 5.97$, $\eta_p^2 = .285$, $p < .05$ respectively[3].

To determine whether this significant overall labeling effect was driven entirely by the match level four condition, i.e., to examine whether there was a matching-features effect at levels one to three, we repeated the previous analyses with match level four excluded. As already noted, there were no significant pairwise differences as a function of match among levels one to three; linear and quadratic trends were also non-significant, $p = .55$ and $p = .19$ respectively. Labeling probabilities pooled across level of match, with match level four excluded, significantly exceeded zero (m = 0.09, $t(33) = 2.26$, $p > .05$), but did not significantly differ from the baseline probability of assigning the same species label to non-matching objects within the same alignable category (m = 0.09, $t(33) = 0.04$, $p = .97$)[4]. Thus, the only significant effect of matching features on labeling in this experiment was at match level four. Identical instances sharing four out of four possible features were given the same species label more often than non-matching but alignable objects; there was no effect of matching features on labeling for non-identical instances, i.e., those sharing from one to three (out of four) matching features.

Visitation Data

The process by which categories are identified is of particular interest in the current experiment, given that our primary research questions pertain to the method underlying category induction, or how categories are discovered. To determine the role of sampling (stage 1) and evaluation (stage 2) in this process, we analyzed the pattern in which other objects were visited while labeling a given object. In particular, we looked at the series of visits made between the time that a person first visits a given object and when they finally enter a label for that object. This series of visits (comparisons to other objects) will be referred to as an *epoch*. Discrete epochs were extracted for each item within the arrays, resulting in 16 epochs for each of the four stimulus displays, or a total of 64 epochs per participant. The first and second repeated instance within each alignable family were coded separately; here we will focus on epochs for the second occurring repeated instance. Analysis of epochs was restricted to those with three to 40 visits between the first visit and category label assignment ($m = 11.47$). Epochs with a length greater than 40 contained more random behavior and hence were less reliable measures of effortful comparisons for a given target object, while epochs shorter than three comparisons contained too few comparisons to be useful.

For regular (non-matching) objects within each alignable category, we compared the probability and proportion of visits to other objects from the same versus different alignable categories.

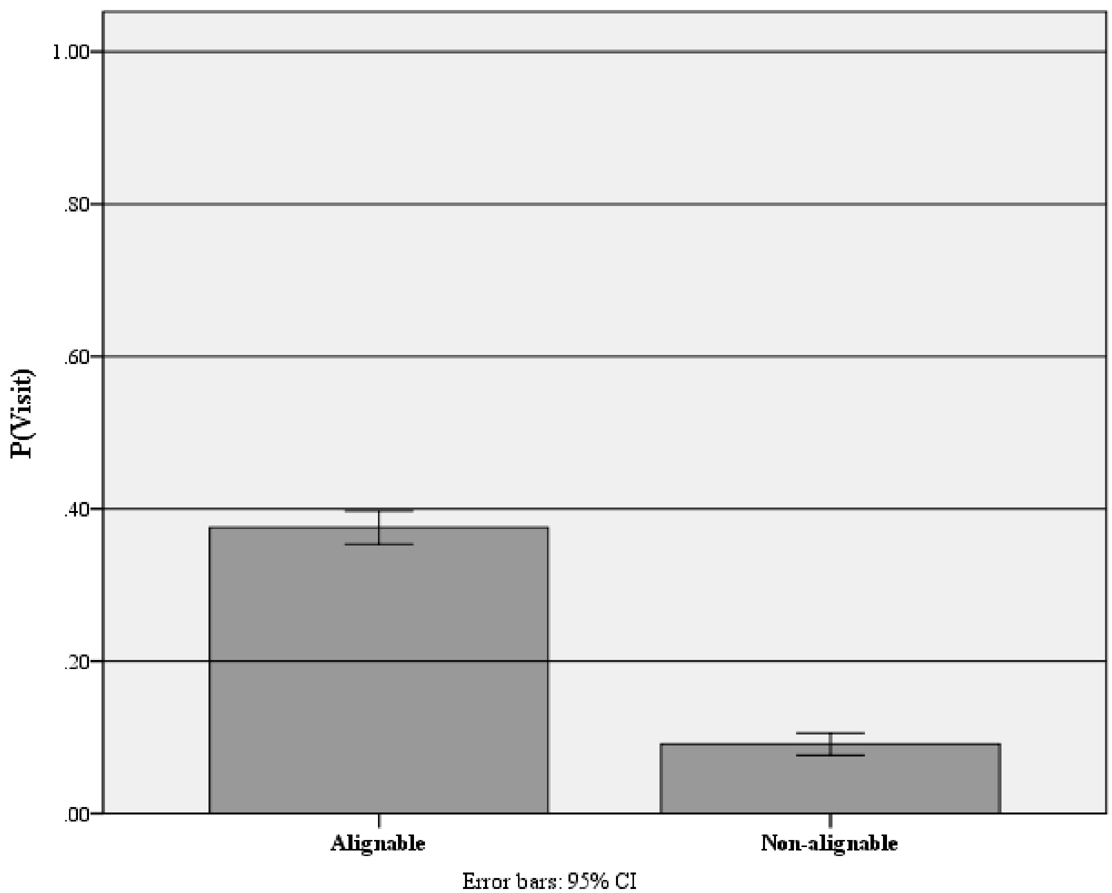


Figure 11. The probability of visiting a given alignable versus non-alignable object at least once during a given epoch.

It seems reasonable to expect that people would be more likely to visit, and show a greater proportion of visits, to other alignable objects in the display compared to non-alignable objects, i.e., that within-category comparisons would be more frequent than between-category comparisons. There was a significant difference in the probability of visiting alignable versus non-alignable objects at least once during an epoch, see Figure 11. The probability of visiting a given alignable object at least once was 0.39, while that of visiting a given non-alignable object was 0.12, $t(36) = 25.38$, $p < .001$. There was also a significant difference between the total proportion of visits to objects from the same versus different alignable categories ($m = 0.08$ vs. 0.04, $t(36) = 8.52$, $p < .001$ (For purposes of comparison, we report the proportion of visits to individual alignable or non-alignable objects, rather than cumulative proportions over all alignable objects of a given type, because there are about twice as many non-alignable as alignable objects in a given display). This result indicates that, in addition to objects from the same alignable category being more likely to receive the same family label, they are also more likely to be visited and received a greater proportion of visits during the process of evaluating category membership.

Our main focus in this experiment was on whether people would be more likely to visit matching than non-matching objects, and whether this would be affected by the level of match (1, 2, 3, 4). First, we report the *probability* of visiting the first matching instance at least once; the *proportion* of visits to matching versus non-matching instances, as well as the average time per visit, will be discussed below. Overall, people were significantly more likely to visit specific matching (m = 0.74) than non-matching but alignable objects (m = 0.37, $t(36) = 6.04$, $p < .001$). For purposes of testing the models, we were most interested in whether the probability of visiting the first of a pair of matching objects at least once when labeling the second was affected by the specific level of match. No effect of level of match on visiting probabilities was observed in this experiment, see Figure 12.

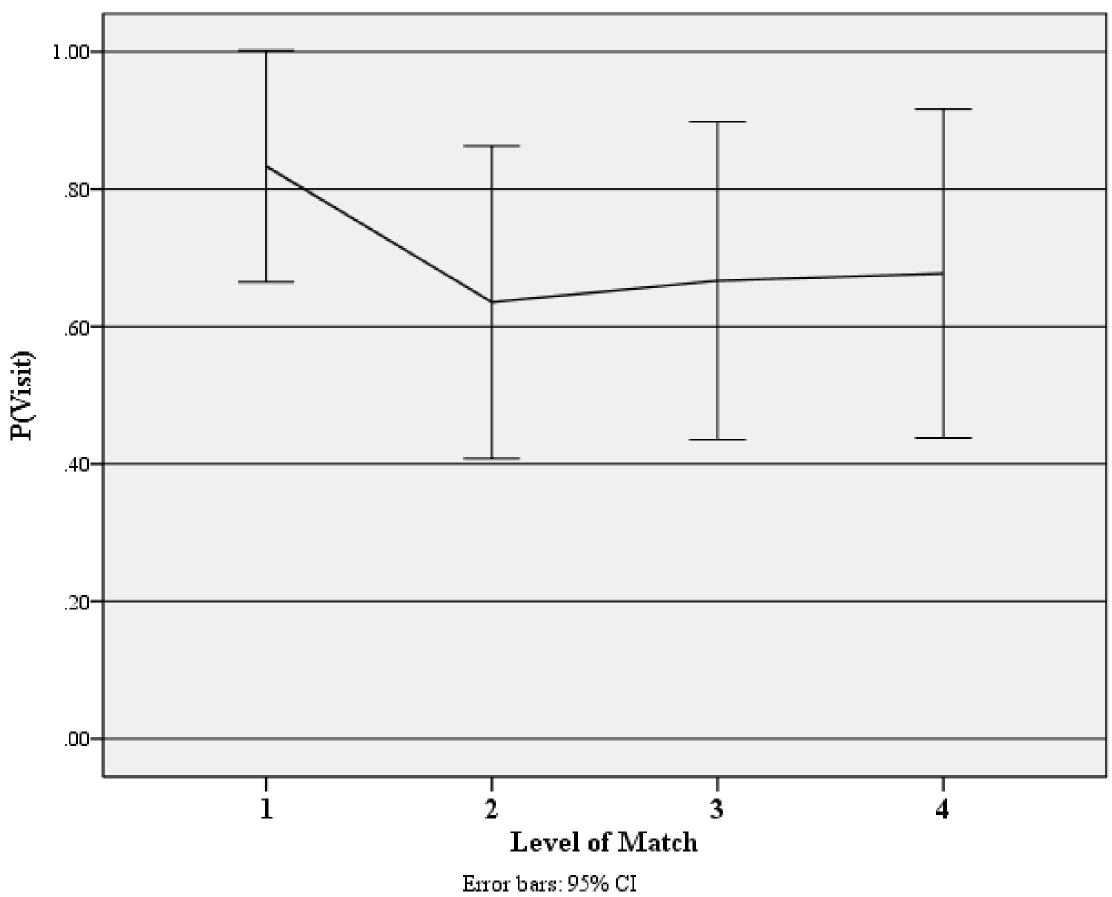


Figure 12. The probability of visiting matching objects as a function of level of match (Second matching objects only).

Objects that shared one feature with the current object (m = 0.83), were just as likely to be visited at least once as objects sharing two (m = 0.64), three (m = 0.67), or four features in common (m = 0.68), $F(3, 45) = 0.94$, $\eta_p^2 = .059$, $p = .43$[3]. Linear and quadratic trend analyses were also not significant, p = .34 and p = .25 respectively. These data provide no evidence that people are more likely to visit objects with more matching features.

To summarize, the probability of visiting a given object was affected by whether it was alignable or shared at least one matching feature with the current object, but there was no further effect of the number of shared features on this measure.

Proportion of Visits

In addition to the probability of visiting an object at least once as a function of level of match, we also examined the overall proportion of visits allocated to the matching object. In principle, the proportion of visits to matching instances could provide additional information about the role of comparing matching objects during the process of categorization. However, the proportion of visits did not vary significantly as a function of match in this experiment (see Figure 13), with identical objects sharing all four features (m = 0.15) being visited just as often as objects sharing one (m = .11), two (m = 0.10) or three features (m = 0.14), $F(3, 45) = 0.86$, $\eta_p^2 = .054$, $p = .47$[3]. Both linear and quadratic trends were non-significant, p = .25 and p = .62, respectively. Averaging across level of match, there was a marginally significant overall difference in visit proportions when matching objects were assigned the same (m = 0.23) versus different (m = 0.16) category label t(13) = 1.85, p = .09[5]. Thus, while there was no effect of level of match on proportion of visits in this experiment, there was some evidence in the data that people allocated a larger proportion of visits to the first matching object when they gave the second object the same label than when they did not.

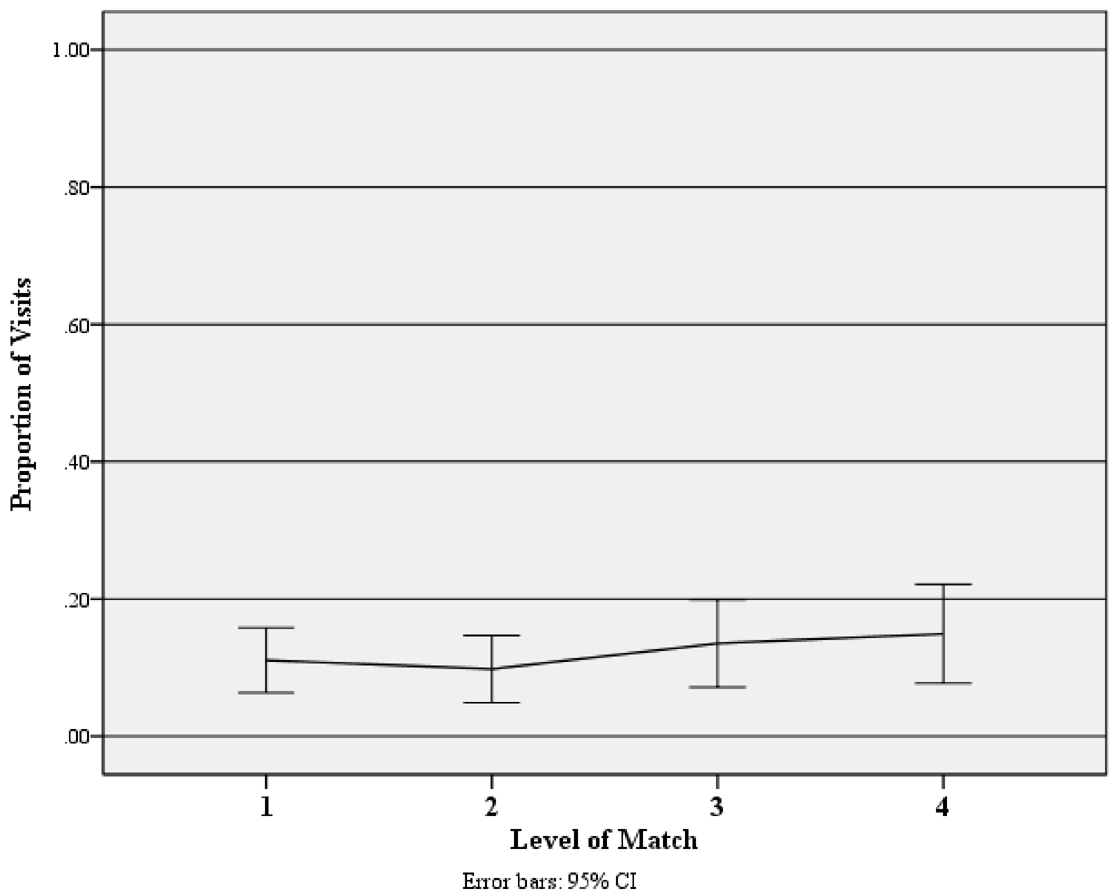


Figure 13. The proportion of visits to matching objects as a function of the level of match.

Viewing Time

As with the proportion of visits to matching instances, the time people spent viewing objects of differing levels of match could provide additional insight into the process of categorization.

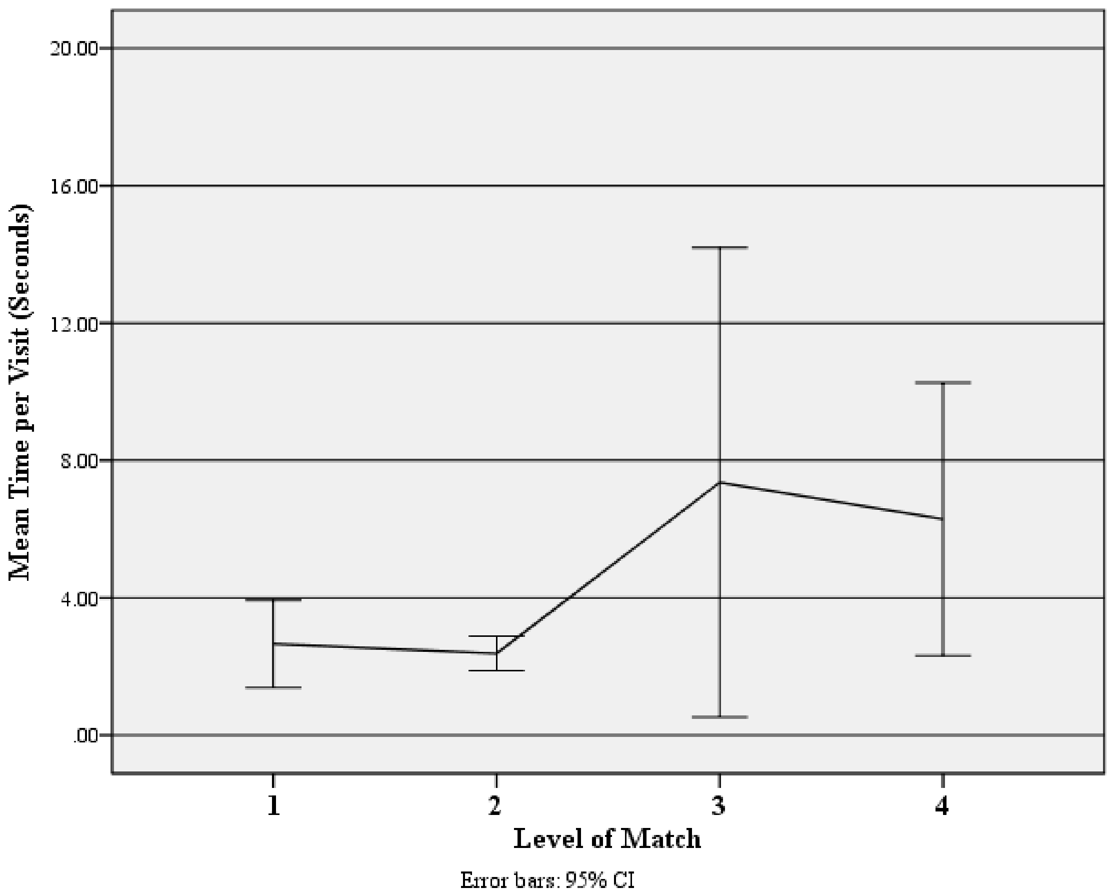


Figure 14. The time spent viewing matching objects as a function of level of match.

The mean viewing time (in seconds) for objects showed a marginally significant effect of level of match, $F(3, 45) = 3.62$, $\eta_p^2 = .194$, $p = .08$; viewing times averaged 6.29 s for objects matching on four features, 7.36 s for those matching on three features, 2.38 s for two features, and 2.66 s for one feature (see Figure 14)[3]. A linear trend analysis was marginally significant $F(1, 15) = 3.62$, $\eta_p^2 = .194$, $p = .08$, though a quadratic trend analyses was non-significant ($p = .75$). There was no significant effect of overall visiting times for matching objects receiving the same (m = 5.53) versus different (m = 4.97) category labels, $t(13) = 0.25$, $p = .81$, averaging over different levels of match[5]. Together, these results suggest that people may have spent somewhat more time visiting objects that matched on three or four features than those that matched on only one or two.

Labels Given Visits

In addition to the probability of giving the second matching object the same label as the first, and the probability of visiting the first object while labeling the second, testing our models also requires computing the conditional probability of the former given the latter, i.e., the conditional probability of assigning the second object the same label as the first given that the first object was visited at least once during the epoch for the second object. Level of match significantly affected this P(same label|visit) measure (see Figure 15), $F(3, 15) = 5.00$, $\eta_p^2 = .500$, $p < .05$. However, when the labeling data is conditionalized in this way only six participants remain with data for each of the four levels of match; this low sample size results in the data having zero variance at match levels 1 to 3. As such, it is more appropriate to compare conditional labeling probability at different levels of match using individual paired-samples t-tests.

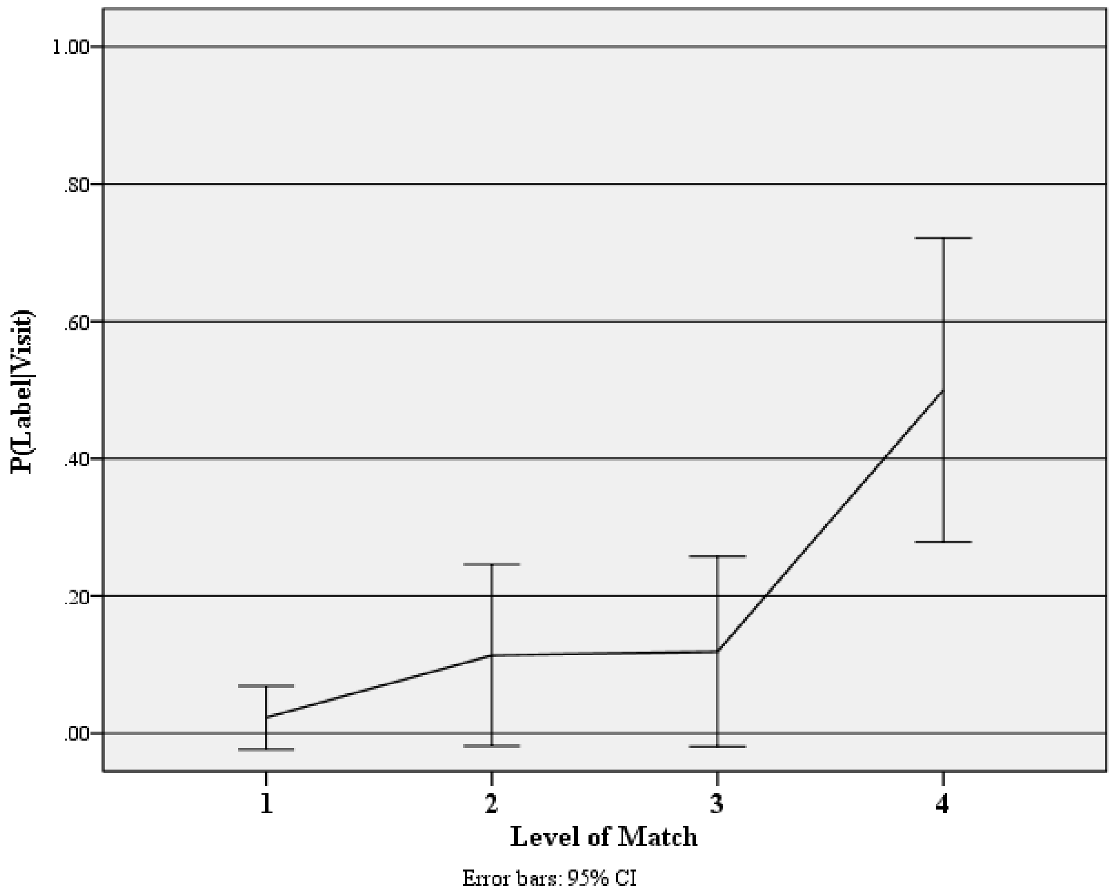


Figure 15. The conditional probability of giving the second matching object the same label as the first, given that the first have been visited at least once, i.e., P(Label|Visit).

This results in a larger N in each comparison and eliminates the problem of zero variance in the lower three levels of match. In these comparisons, identical instances sharing all four features were assigned the same species label significantly more often ($m = 0.46$) than objects sharing one ($m = 0.0$, $t(12) = 3.21$, $p < .01$), two ($m = 0.08$, $t(12) = 2.74$, $p < .05$), or three features ($m = 0.08$, $t(11) = 2.35$, $p < .05$), conditional on visiting the matching object at least once during that epoch[6]. There were no significant differences among the later three groups, and none significantly exceeded zero. Thus, as with the non-conditionalized labeling data, the main difference was between identical pairs and all other levels of match.

CHAPTER FOUR

GENERAL DISCUSSION

Summary of Results

The current experiment demonstrates sensitivity to alignable category structure, with objects of the same structural configuration, or body plan, being assigned the same family and species labels considerably more often than non-alignable objects differing in structural configuration, despite sharing no identical features in common. In addition, matching objects within alignable categories were assigned the same species label more often than other alignable objects that shared no identical features in common. However, this was entirely due to match level four being assigned the same species label more often than other levels of match, which did not differ among themselves or from non-matching objects within the same alignable families. In general, these results are consistent with previous results in showing sensitivity to alignability and matching features. However, rather than showing a graded similarity effect across levels of match, as in previous studies, the labeling data from the current experiment only shows an effect of complete identity versus non-identity within alignable categories.

The probability of visiting prior objects was affected by whether the objects was alignable with the current object, and also by whether they shared any matching features. However, it was unaffected by number of matching features shared by the two objects. In other words, the probability-of-visit data shows a pattern of non-random scanning over prior objects, but no effect of level of match. We also assessed the average duration of visits and the proportion of visits to matching objects as a function of level of match. The results indicated no significant variation in the proportion or average duration of visits to different objects as a function of level of match (although the latter result did attain marginal significance).

To assess our models, we also computed the conditional probability of assigning the same category label to matching objects given that the first matching object was visited at least once during the epoch for the second matching instance. This calculation allows for a pure examination of the comparison-decision process (stage 2) on category outcomes by removing the sampling probability from the labeling data. The probability of grouping matching instances together given that they were visited was significantly affected by the number of features shared by the two matching objects, with P(label|visit) increasing with a larger number of shared features. However, as with the non-conditionalized labeling data, only match level four was significantly greater than lower levels of match and there were no differences between match levels one, two, and three.

Assessing the Models

While level of match affected the probability of same-category label assignment, as in previous experiments, the probability of visiting the first matching objects at least once (stage 1) while labeling the second was unaffected by the number of features they shared. This suggests that there is no effect of match level on the stage 1 sampling probability in this task. When labeling is conditionalized on visiting matching objects, thus eliminating the effect of visiting probability on the labeling data, there is a strong effect of match level that lends support to the stage-2-only model. In other words, a participant who visits the first matching instance is more likely to give the second the same label if they share more features, but sharing more features does not increase the probability of such a visit occurring in the first place. Given that the level of match has no effect on sampling probabilities, the stage 2 only model, in which a comparison and evaluation of category membership is the primary predictor of same-category assignment, is best supported by the present data.

Although the present data provide no evidence for a stage 1 contribution to the levels-of-match effect, they are consistent with an independent effect of stage 1 in categorization more generally. Importantly, the visitation measure was affected by some of our independent variables in this experiment. People were more likely to visit alignable objects, as well as those with matching features, than non-alignable and non-matching objects. People could have shown a greater one-time probability of visiting alignable objects not only by remembering which previously visited positions in the display contained objects that were alignable with the current object, but the blurred silhouettes of the objects in the display would also have allowed the person to tell to some extent what kind of object each was. On the other hand, previous feature matches would have to be remembered in order to have an effect on initial visit probability, as no information about specific matching features could be gleaned from the obscured objects.

It is interesting that people were more likely to visit matching instances that shared a single feature, with additional features making no further contribution to this effect, while only a match level of four was sufficient to have an actual effect on the probability of categorization. This is in sharp contrast to some of the sorting experiments reviewed in the Introduction (e.g., Medin et al., 1987), in which people constructed categories based on a single feature while ignoring all other features of the objects. This difference may be due in part to the unforced nature of the present task, in which people were under no pressure to put the matching instances together into the same species category. The higher variability (larger number of values per dimension) in the present experiment may also have contributed to this difference, given that dividing a set of objects into dichotomous categories along a binary dimension seems such an obvious strategy in a sorting task.

Problems, Issues, and Limitations

One difference between the present data and prior results using the binomial labeling task is that there is a somewhat smaller effect of feature match in the current experiment. Only the labeling probability at match level four was significantly greater than that for non-matching but alignable objects. Thus, there was an overall *identity* effect in the present experiment, but not a graded *similarity* effect as in the previous studies. One possible reason for this difference has to do with the nature of the present scanning task. For example, the stimulus presentation utilized in the current design is neither truly simultaneous nor sequential, as typically defined in previous categorization research. A simultaneous display enables participants to view all objects at once and directly compare objects with relative ease. In a sequential task, participants usually see one object at a time and the order of presentation is controlled by the experimenter, not the participant. The current design requires that participants navigate an array sampling objects one at a time, but allows them to view these objects in any order they want, or scan back and forth between particular objects at will, much as they could in a simultaneous condition. However, the fact that objects can only be seen one at a time may make it difficult to track individual matching features in this task, and thus account for the decrease in labeling probabilities at lower levels of match compared to those obtained in previous full-set paper-and-pencil studies. This suggests that the method by which objects are displayed may have an effect on participants' sensitivity to similarity and the shape of the similarity function. Comparing the present results to those obtained with a full-set simultaneous display used in conjunction with an eye tracker, or a truly sequential paradigm in which people look at each only object once, may provide additional insight into the sampling and evaluative processes underlying free categorization.

The lack of a stage 1 sampling effect as a function of match is also a somewhat unexpected result, especially given the significant effects of both alignability and match versus mismatch on our probability-of-visit measure, and it is unclear whether it will generalize to other presentation paradigms. This lack of effect does not appear to be due to a mere lack of statistical power, as there is no suggestion of any trend in the probability of visiting matching instances as a function of level of match in the present data (by contrast, some of the other level-of-match effects did attain statistical significance in this data). One possible reason for this lack of an effect is the fact that people can only see one example at a time in the current task, so prior objects can only capture the person's current attention via memory reminding. This contrasts with a full-set presentation in which two objects sharing a salient distinctive feature might simultaneously appear to *pop out* of the stimulus display (e.g., two red X's might appear to pop out from a field of green X's). In principle, this concern could be addressed by using simultaneous stimulus presentation in conjunction with eye tracking technology, which would provide measures of sampling and evaluation without the somewhat artificial task of navigating an array of obscured objects while only being able to examine one object at a time. Another possibility would be to pre-expose participants to the entire array, allowing them to view objects simultaneously and freely compare them, before beginning the current scan-and-label task. People might show a less random sampling pattern following such pre-exposure.

Larger Implications

The current experiment was designed to investigate factors associated with the basic act of freely assigning two stimuli to their own newly-created category, which is arguably the essence of free categorization – or at least its essential first step. We have proposed that free categorization depends on two factors, an initial sampling of objects to compare (stage 1) and a subsequent evaluation to determine if their similarity warrants placing them into the same category (stage 2). These two stages can in principle be affected by different independent variables, or affected differentially by the same independent variables. The two stage model of categorization proposed here resembles models of analogy (e.g., Gentner & Forbus, 2011) in which there is an initial retrieval stage, during which a target and source may be compared, followed by a mapping stage during which it is determined whether an actual analogy can be constructed. The retrieval stage is largely affected by obvious surface features, while the mapping stage is determined by the abstract alignability of the objects being compared.

The current results did not provide support for stage 1 as a distinct contributor to the level-of-match effect in the present task. However, the data do show that people are more likely to sample prior alignable objects than non-alignable objects while labeling a current object, as well as objects that share at least one matching feature. They also allocate a marginally larger proportion of visits to the first matching object when giving the second the same species label than when they are labeled differently. These results suggest that the current task may prove useful in further research aiming to assess the independent contribution of stages 1 and 2 to free categorization in a various situations. There are likely additional factors that influence the stage 1 sampling probability, such as spatial proximity in a display or perceptual salience of shared individual features, that can be manipulated to investigate this stage. As such, we expect the task used in the current experiment and its variants (e.g., in which order of presentation is made an independent rather than dependent variable) to continue to be a useful tool for investigating fundamental issues of free categorization.

APPENDIX A

TASK INSTRUCTIONS

INSTRUCTIONS

In this experiment, we'd like you to imagine yourself in the role of an inter-planetary biologist who has just received a shipment of fossilized Martian organisms. These creatures are completely new to science, and your job is to classify them – to sort them into broad families, as well as to identify individual species.

To be clear what we mean, a biological family refers to some broad group like cats, fish, trees, or birds. Individual species refer to specific kinds within a broader family. For example, lion, tigers, and leopards are species of cats; trout, tuna, and sharks are species of fish; oaks, maples, and pines are species of trees; and crows, finches, and sparrows are species of birds.

Family	Species
Cats	Lions, Tigers, Leopards
Fish	Trout, Tuna, Sharks
Trees	Oaks, Maples, Pines
Birds	Crows, Finches, Sparrows

In this study, you will examine six pictures of 16 specimens (supposed fossilized organisms), arranged in no particular pattern within a 4x4 grid. Each creature will be clearly visible one at a time using the mouse to navigate to the labeling area below each creature. Your job will be to divide these creatures into families and species. Specifically, you will be labeling the creatures by family (A, B, C, etc.) and by species within each family (1, 2, 3, etc.). Feel free to label the creatures in any order, and to go back to adjust your labels prior to moving on to the next picture of creatures.

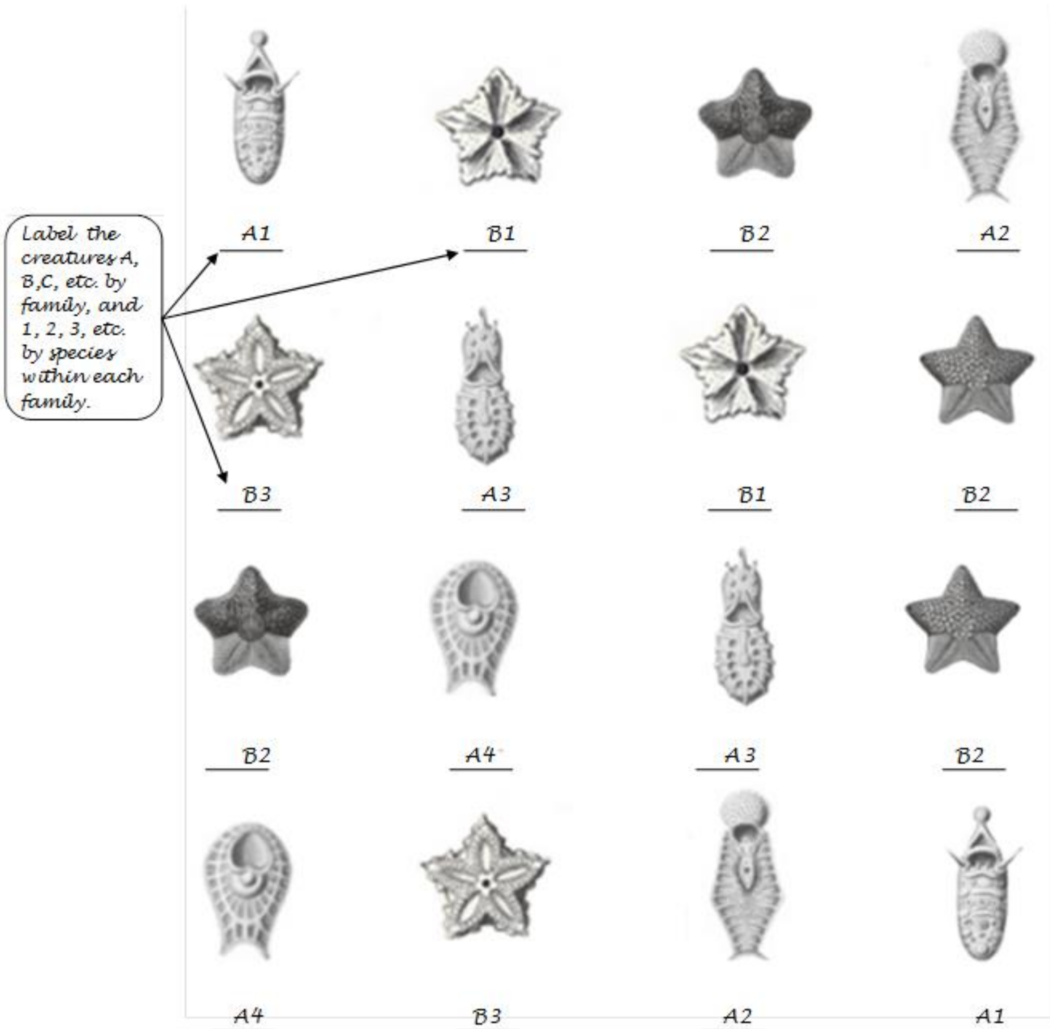
The sample screen following these instructions will show you how to do this.

When you are ready to view the sample, press the SPACEBAR

APPENDIX B

SAMPLE DISPLAY

SAMPLE PAGE

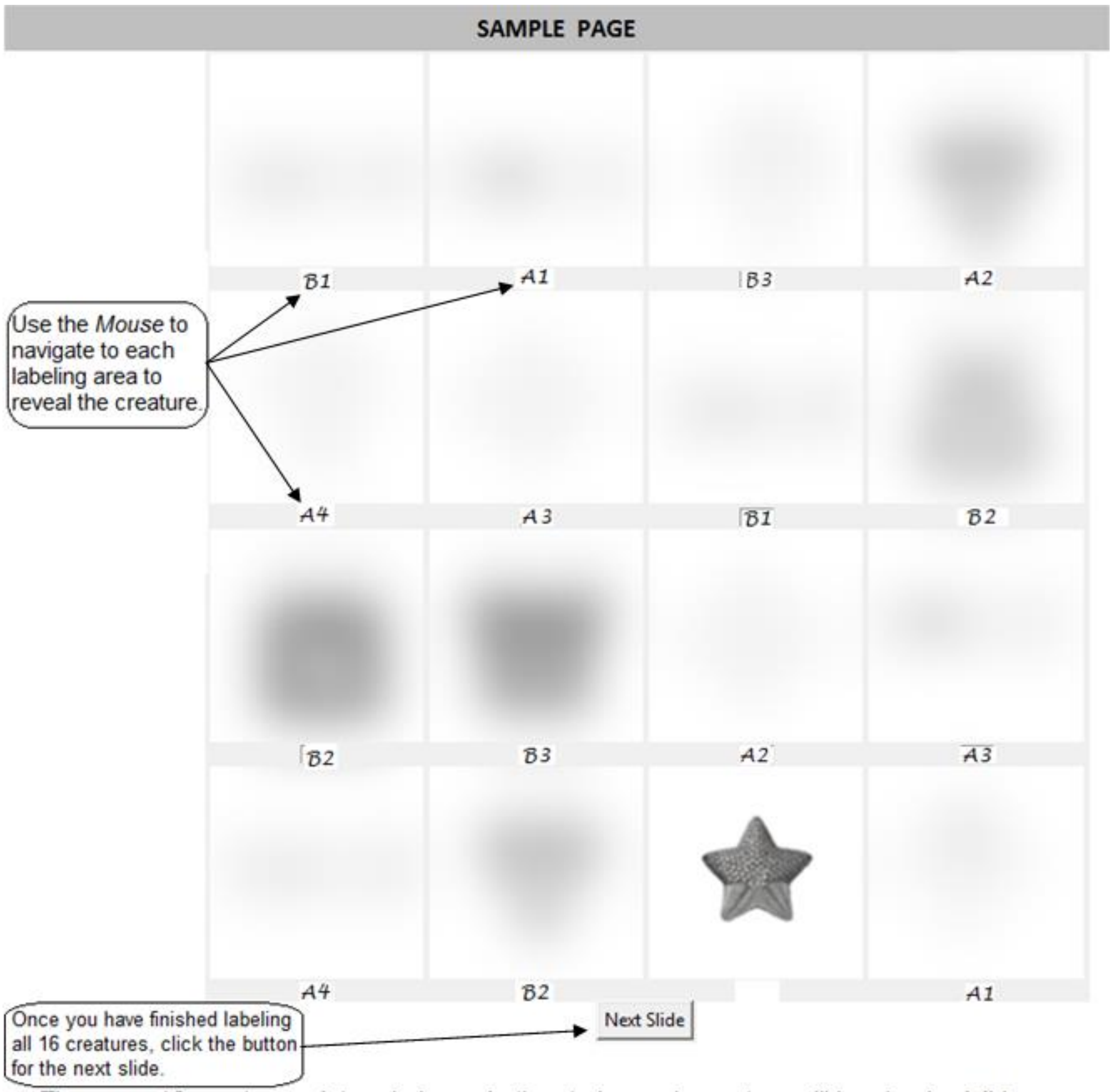


There are 16 creatures pictured above. In the study, each creature will be clearly visible one at a time using the mouse to navigate to the labeling area below each creature. Label each creature A, B, C, etc. by family/group and 1, 2, 3 etc. by individual species within each family. Feel free to label the creatures in any order, and to go back to adjust your labels prior to moving on to the next picture of creatures.

Press the SPACEBAR to Continue
to the Next Sample

APPENDIX C

SAMPLE TASK DISPLAY

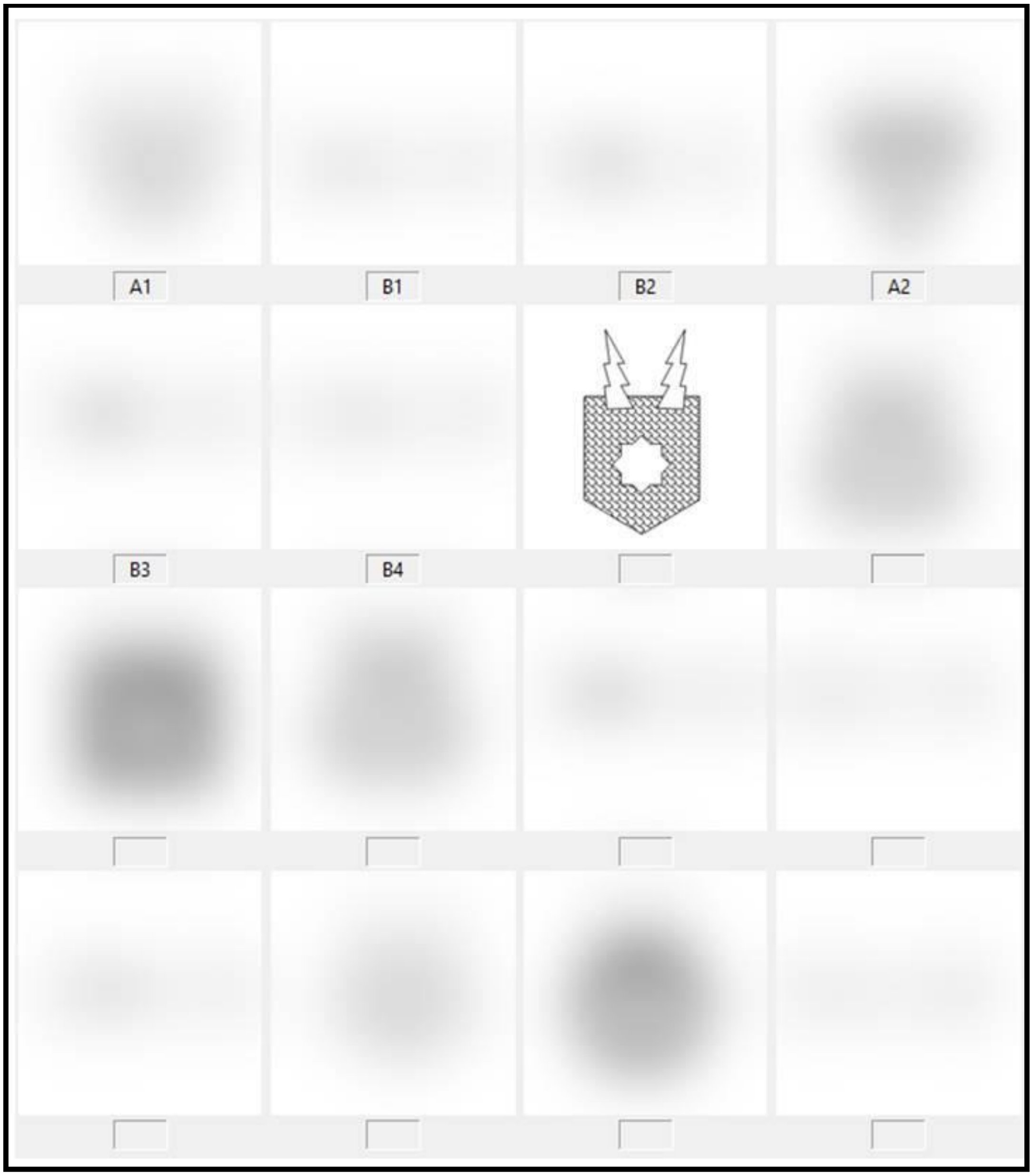


There are 16 creatures pictured above. In the study, each creature will be clearly visible one at a time using the mouse to navigate to the labeling area below each creature. Label each creature A, B, C, etc. by family/group and 1, 2, 3 etc. by individual species within each family. Feel free to label the creatures in any order, and to go back to adjust your labels prior to moving on to the next picture of creatures.

Press the SPACEBAR to Begin

APPENDIX D

TASK DISPLAY


A1
B1
B2
A2
B3
B4

APPENDIX E

IRB APPROVAL

Human Subjects Review Board
Department of Psychology
California State University,
San Bernardino

PI: Gregory Smith and John Clapper

From: Jason Reimer

Project Title: Comparison and Categorization in Visual Arrays

Project ID: H-15WI-17

Date: 2/24/2015

Disposition: Expedited Review

Your IRB proposal is approved. This approval is valid until 2/24/2016.

Good luck with your research!

Jason Reimer, Co-Chair
Psychology IRB Sub-Committee

REFERENCES

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16*, 81-121.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369-406.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

Clapper, J. P. (in press). The impact of training sequence and between-category similarity on unsupervised induction. *The Quarterly Journal of Experimental Psychology.*

Clapper, J. P. (2015). *Investigations of free categorization using a binomial labeling task.* Unpublished manuscript, CSUSB.

Clapper, J. P., Smith, G. J., & Miller, B. A. (2015). *Similarity-based free categorization in a binomial labeling task.* Unpublished manuscript, CSUSB.

Corter, J., & Gluck, M. (1985). Explaining basic categories: Feature predictability and information. *Psychological Bulletin, 111*, 291-303.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley Interdisciplinary Reviews-Cognitive Science*.

Haeckel, E. (1974). *Art Forms in Nature*. New York: Dover Publications.

Handel, S., & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception & Psychophysics, 12*, 108–116.

Imai, S., & Garner, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology, 69*, 596-608.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

Mervis, C. R., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Milton, F. N., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 407–415.

Milton, F. N., Longmore, C. A., & Wills, A. J. (2008). Processes of Overall Similarity Sorting in Free Classification. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 676–692.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303-343.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., &

McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition, 121*, 83–100.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 347-363.

Rosch, E. H., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E. H., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 491-502.

Smith, J. D., & Kemler Nelson, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General, 113*, 137–159.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-350.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113*, 169-193.

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception*. Journal of Experimental Psychology: Human Perception and Performance, 9*, 103–112.

Footnotes

[1]The degrees of freedom are reduced in the ANOVA as only 16 of 46 participants had data for all four levels of match. This is largely due to missing labeling data as a result of restricting the range of epoch lengths included in our analysis. Despite the reduced sample size, our effects on this measure remained statistically significant.

[2]There are a total of 20 cases in which participants had data on all three levels of match (1, 2, and 3), which accounts for the reduction in degrees of freedom in this analysis, hence df = 38.

[3]Again, this analysis assumes a sample size of 16 participants with data at all four levels of match.

[4]When data is pooled across match level one, two, and three, there are a total of 34 cases in which participants had data on at least one of these three levels of match, hence df = 33.

[5]There are a total of 14 cases in which matching objects were assigned both the same and different species labels by the same participant, reducing the degrees of freedom to 13 for these analyses.

[6]The degrees of freedom vary slightly in these t-tests because the data have different numbers of missing cells across different levels of match.